Some contributions to model selection and statistical inference in Markovian models

Shouto Yonekura

Submitted for the degree of Doctor of Philosophy at the Department of Statistical Science, UCL.

30th August 2020

Declaration

I, Shouto Yonekura, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

The general theme of this thesis is providing and studying a new understanding of some statistical models and computational methods based on a Markov process/chain. Section 1-4 are devoted to reviewing the literature for the sake of completeness and the better understanding of Section 5-7 that are our original studies.

Section 1 is devoted to understanding a Markov process since continuous and discrete types of a Markov process are hinges of the thesis. In particular, we will study some basics/advanced results of Markov chains and Ito diffusions. Ergodic properties of these processes are also documented.

In Section 2 we first study the Metropolis-Hastings algorithm since this is basic of other MCMC methods. We then study more advanced methods such as Reversible Jump MCMC, Metropolis-adjusted Langevin algorithm, pseudo marginal MCMC and Hamiltonian Monte Carlo. These MCMC methods will appear in Section 3, 4 and 7.

In Section 3 we consider another type of Monte Carlo method called sequential Monte Carlo (SMC). Unlike MCMC methods, SMC methods often give us on-line ways to approximate intractable objects. Therefore, these methods are particularly useful when one needs to play around with models with scalable computational costs. Some mathematical analysis of SMC also can be found. These SMC methods will appear in Section 4, 5, 6 and 7.

In Section 4 we first discuss hidden Markov models (HMMs) since all statistical models that we consider in the thesis can be treated as HMMs or their generalisation. Since, in general, HMMs involve intractable objects, we then study approximation ways for them based on SMC methods. Statistical inference for HMMs is also considered. These topics will appear in Section 5, 6 and 7.

Section 5 is largely based on a submitted paper titled *Asymptotic Analysis of Model Selection Criteria for General Hidden Markov Models* with Alexandros Beskos and Sumeetpal Sidhu Singh, https://arxiv.org/abs/1811.11834v3. In this section, we study the asymptotic behaviour of some information criteria in the context of hidden Markov models, or state space models. In particular, we prove the strong consistency of BIC and evidence for general HMMs.

Section 6 is largely based on a submitted paper titled *Online Smoothing for Diffusion Processes Observed with Noise* with Alexandros Beskos, https://arxiv.org/abs/2003.12247. In this section, we develop sequential Monte Carlo methods to estimate parameters of (jump) diffusion models.

Section 7 is largely based on an ongoing paper titled *Adaptive Bayesian Model Selection for Diffusion Models* with Alexandros Beskos. In this section, we develop adaptive computational ways, based on sequential Monte Carlo samplers and Hamiltonian Monte Carlo on a functional space, for Bayesian model selection.

Impact statement

The work presented in this thesis has a potential impact on both academic and industrial communities.

First, the studies shed light on a new understanding of model selection methods in the context of hidden Markov models. Indeed, we have verified model selection consistency which has been an open problem in the context of hidden Markov models. Since model selection problems also frequently arise in practice, our theoretical results could be useful guidance for industrial communities. Here we emphasise that hidden Markov models are routinely used in such diverse disciplines as finance, speech recognition and epidemiology so that they are practical statistical models.

Also, this thesis provides novel algorithms based on sequential Monte Carlo in the context of statistical inference for (jump) diffusion models. In particular, we develop algorithms on the infinite-dimensional path-space under the weak assumptions commonly used in the literature. Therefore, our algorithms are applicable to a wide class of diffusion models compared with the literature. This study is of great importance for industrial communities as well since the implementation of our algorithms is quite simple so that non-experts could easily use them. It should be emphasised that diffusion models are extremely popular ones among financial companies.

Finally, this study also explores ways to make use of SMC samplers in the context of Bayesian model selection. In particular, we again focus on diffusion models. Research on Bayesian model selection for diffusion models has been limited despite its great importance for both academic and industrial communities. The main difficulty comes from the high-dimensional nature of diffusion models. Our study could be a guiding principle to deal with this problem. Also, we show that SMC samplers give rise to natural methods to learn adaptively tuning parameters.

Acknowledgements

First and foremost, I would like to sincerely express my deepest gratitude to my supervisor Alexandros Beskos for his encouragement, humour, patience and teaching me exciting research topics. I feel privileged to have studied under his supervision. Special thanks are needed for Sumeetpal Sidhu Singh, Yvo Pokern, Samuel Livingstone, Ricardo Silva and Theodoros Damoulas. I am also grateful to Hisashi Tanizaki and Kengo Kamatani who helped and encouraged me to study at UCL. Finally, I acknowledge funding from the Alan Turing Institute under grant number TU/C/000013. Any remaining inaccuracies are, of course, because of the author.

No words could adequately express my gratitude to my parents and partner for their all dedicated support. Without them, this thesis would not have existed.

Contents

List of Algorithms

List of Figures

Frequently used notations

- Let $(E, \mathcal{E})$ be a measurable space. For a function $f : E \to \mathbb{R}$, we write $\|f\|_\infty := \sup_{x \in E} \mid f(x) \mid$, and we denote the set of all functions such that $\|f\|_\infty < \infty$ by $\mathcal{B}_b(E)$.
- Given a measurable space $(E, \mathcal{E})$, we denote the set of all finite signed measures by $\mathcal{S}(E)$, the set of all probability measures by $\mathcal{P}(E) \subset \mathcal{S}(E)$.
- For some $\mu(dx) \in \mathcal{P}(E)$, $x \overset{i.i.d.}{\sim} \mu(dx)$ denotes that a random variable x is independent and identically distributed according to $\mu(dx)$.
- Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. We write $\mathbb{E}$ to denote the expectation with respect to $\mathbb{P}$, and the variance of a random variable X is denoted as $\mathbb{V}[X] := \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$. The covariance of random variables X and Y is denoted by $\mathrm{Cov}(X, Y) := \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$.
- For $p \geq 1$ and given a measurable space $(E, \mathcal{E}, \mu)$, $L^p(\mu)$ denotes the set of $\mu-$equivalent functions in the set such that $\left\{ f : E \to \mathbb{R}^d; \left(\int_E \mid f(x) \mid^p \mu(dx) \right)^{1/p} < \infty \right\}$.
- For any $x, y \in \mathbb{R}^d$, $\langle x, y \rangle$ denotes $\sum_{i=1}^d x_i y_i$ and $\|x\|$ denotes $\sqrt{\sum_{i=1}^d x_i^2}$.
- $\mathcal{N}(\mu, \Sigma)$ denotes the Gaussian distribution on $\mathbb{R}^d$ with mean $\mu \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$. Also, $\mathcal{N}(x; \mu, \Sigma)$ denotes its density $x \mapsto \mathcal{N}(x; \mu, \Sigma)$.
- We write $f(n) = \mathcal{O}(g(n))$ if $\lim_{n \to \infty} \frac{f(n)}{\mathcal{O}(g(n))} < \infty$ holds.
- We use w.r.t. as an abbreviation for with respect to, w.p.1 as an abbreviation for with probability 1, and iff as an abbreviation for if and only if.

1 Markov processes

1.1 Introduction

The statistical models and algorithms will be introduced and developed in this thesis are mainly based on a Markov process, which is a special class of a stochastic process. Therefore, this chapter is devoted to providing some basic and advanced results of Markov processes. In particular, we will focus on the two important Markov processes, that is Markov chains and Ito diffusions.

1.2 General properties of a Markov process

1.2.1 Basics of stochastic processes

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(E, \mathcal{E})$ be a measurable space. Following Pavliotis (2014); Çınlar (2011) closely, we first begin with the definition of a stochastic process.

Definition 1. Let $(E, \mathcal{E})$ be a measurable space and T be an arbitrary set. For each $t \in T$, denote by $X_t(\omega)$ a random variable taking values in $(E, \mathcal{E})$. Then, the collection $\{X_t : t \in T\}$ is called a *stochastic process* with a *state space* $(E, \mathcal{E})$ and a *parameter set* T.

In general, one might take $E = \mathbb{R}^d$ with $d \geq 1$, and $\mathcal{E} = \mathcal{B}(\mathbb{R}^d)$. When $T = \mathbb{Z}_+$, then a stochastic process $X_t(\omega)$ is called a discrete time stochastic process, and when $T = \mathbb{R}_+$, $X_t(\omega)$ is called a continuous time stochastic process. In the sequel of this section, we will focus on the case when $T = \mathbb{R}_+$ and $E = \mathbb{R}^d$.

Clearly, a stochastic process $X_t(\omega)$ has two inputs, that is $t \in T$ and $\omega \in \Omega$. For given $\omega \in \Omega$, the map $t \mapsto X_t(\omega)$ is called a *sample path* or *trajectory* of the process X. Also, given $t \in T$, the map $\omega \mapsto X_t(\omega)$ is a random variable. We will denote the map $t \mapsto X_t(\omega)$ by X_t and the map $\omega \mapsto X_t(\omega)$ by X. Notice that the process $\{X_t : t \in T\}$ can be considered as a random variable X which takes values in the product space $(E^T, \mathcal{E}^{\otimes T})$. The distribution (or law) of the random variable X on the product space $(E^T, \mathcal{E}^{\otimes T})$ is then given as the push-forward measure, which might be expressed as $Law(X) := \mathbb{P}(X^{-1}(A))$ for $A \in \mathcal{E}^{\otimes T}$.

Apparently, one can seldom describe the law of X explicitly. Recall that the product $\sigma-$algebra $\mathcal{E}^{\otimes T}$ is generated by the finite dimensional rectangles. Therefore, a probability measure on $(E^T, \mathcal{E}^{\otimes T})$ can be identified by the following *finite dimensional distributions.*

Definition 2. Let $t_i \in T$ and $i \in \{1, 2, \cdots k\}$. The *finite dimensional distributions (*FDDs*)* of a stochastic process $\{X_t : t \in T\}$ are defined by:

$$\mathbb{P}\left(X_{t_1} \in A_1, X_{t_2} \in A_2, \cdots, X_{t_k} \in A_k\right),$$

where $A_i \in \mathcal{E}$ for each $i \in \{1, 2, \cdots k\}$.

In general, we are particularly interested in long time behaviour of a stochastic process $\{X_t : t \in T\}$. Then it is *stationarity* to characterise such long tine behaviour. A process satisfies the following definition is often said to be a *strongly stationary process* or a *strictly stationary process.*

Definition 3. A stochastic process $\{X_t : t \in T\}$ is said to be *strongly stationary* if

$$\mathbb{P}\left(X_{t_1} \in A_1, X_{t_2} \in A_2, \cdots, X_{t_k} \in A_k\right) = \mathbb{P}\left(X_{t_1+s} \in A_1, X_{t_2+s} \in A_2, \cdots, X_{t_k+s} \in A_k\right),$$

holds for any $s \in T$.

Loosely speaking, definition 3 requires that all FDDs of process $\{X_t : t \in T\}$ are invariant under time shift for any integer k and time time parameter t. Clearly, strongly stationary processes are *identically distributed* for all t. Assume that $\mathbb{E}\left[X_t^2\right] < \infty$ for any t, that is $X_t \in L^2(\mathbb{P})$. Then notice that if $\{X_t : t \in T\}$ is strongly stationary, we have that:

$$\begin{aligned}\mathbb{E}\left[X_{t+s}\right] &= \mathbb{E}\left[X_t\right], \\ \operatorname{Cov}\left(X_t, X_{t+s}\right) &= \operatorname{Cov}\left(X_1, X_{1+s}\right),\end{aligned}$$

for all $s \in T$. Therefore, we can observe that the mean of a strongly stationary process is constant and the covariance of it only depends on the difference between the two time parameters t and s. These observations lead us to the following definition.

Definition 4. Assume that $\mathbb{E}\left[X_t^2\right] < \infty$ for any t. A stochastic process $\{X_t : t \in T\}$ is said to be a *weakly stationary* (process) if the followings hold:

i) $\mathbb{E}\left[X_t\right] = \mu$ for any $t \in T$.

ii) $\gamma(t, s) := \mathbb{E}\left[\left(X_t - \mu\right)\left(X_s - \mu\right)\right] = \gamma(t - s)$ for any $t, s \in T$.

As we have studied, a strongly stationary process (with finite second moment) is a weakly stationary process but not vice versa. For instance, let $Z_t \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$ and set:

$$X_t := \begin{cases} Z_t & if\, t\, is\, even, \\ \frac{1}{\sqrt{2}}(Z_{t-1}^2 - 1) & if\, t\, is\, odd. \end{cases}$$

Then one can show that $\mathbb{E}\left[Z_t\right] = 0$ and $\mathbb{V}\left[X_t\right] = 1$ for both even and odd t, and hence $\operatorname{Cov}(X_t, X_{t+s}) = 0$. Thus X_t is weakly stationary. Whilst $\mathbb{P}(X_t < x_t) = \mathbb{P}(Z_t < x_t)$ for all even t, we have that:

$$\begin{aligned}\mathbb{P}(X_t < x_t) &= \mathbb{P}\left(\frac{1}{\sqrt{2}}(Z_{t-1}^2 - 1) < x_t\right), \\ &= \mathbb{P}\left(-\sqrt{\sqrt{2}x_t + 1} < Z_{t-1} < \sqrt{\sqrt{2}x_t + 1}\right),\end{aligned}$$

for all odd t. Clearly $\mathbb{P}(X_t < 0) = 0.5$ for all even t and $\mathbb{P}(X_t < 0) \neq 0.5$ for all odd t. Thus such X_t is not strongly stationary in this case. In the sequel, we will use stationary for strong stationary sense.

Definition 5. Let $(E, \mathcal{E})$ and $(F, \mathcal{F})$ be measurable spaces. The mapping $K : E \times \mathcal{F} \to \mathbb{R}_+$ is called a *transition kernel* from $(E, \mathcal{E})$ into $(F, \mathcal{F})$ if K satisfies the followings:

i) The mapping $x \mapsto K(x, B)$ is $\mathcal{E}-$measurable for any $B \in \mathcal{F}$.

ii) The mapping $B \mapsto K(x, B)$ is a measure on $(F, \mathcal{F})$ for any $x \in E$.

If $K(x, F) = 1$ for all $x \in E$, then K is called a *probability transition kernel*. For $f \in \mathcal{B}_b(F)$, importantly, a transition kernel K induces the following operators so-called *right Multiplication* and *left Multiplication,* see, e.g, Çınlar (2011, Theorem 1.6.3):

$$Kf(x) := \int_F K(x, dy) f(y), \tag{1.1}$$

$$\mu K(B) := \int_E \mu(dx) K(x, B). \tag{1.2}$$

Proposition 1. *Let $f \in \mathcal{B}_b(F)$ and K be a transition kernel from $(E, \mathcal{E})$ into $(F, \mathcal{F})$. Then for any $x \in E$, $Kf(x) = \int_F K(x, dy) f(y) \in \mathcal{B}_b(E)$. Also let $\mu \in \mathcal{S}(E)$. Then for any $B \in \mathcal{F}$, $\mu K(B) = \int_E \mu(dx) K(x, dy) \in \mathcal{S}(F)$. Also, for $\mu \in \mathcal{S}(E)$ and $f \in \mathcal{B}_b(F)$, $(\mu K) f = \mu(Kf) = \int_E \mu(dx) \int_F K(x, dy) f(y)$.*

1.2.2 Markov process

We will introduce a special class of a stochastic possess. Let $\mathcal{F}_t$ be a filtration generated by a stochastic process X_t, that is, a non-decreasing family of the smallest $\sigma-$ algebra such that X_t is a measurable function w.r.t. it. Using a filtration $\mathcal{F}_t$, we can define the following particularly important class of a stochastic process.

Definition 6. (Markov process). A stochastic process $\{X_t : t \in T\}$ on a measurable space $(E, \mathcal{E})$ is said to be a *Markov process* if for all $s, t \in T$ with $s < t$ and $B \in \mathcal{E}$:

$$\mathbb{P}(X_t \in B \mid \mathcal{F}_s) = \mathbb{P}(X_t \in B \mid X_s) \tag{1.3}$$

holds, where $\mathcal{F}_t$ is the filtration generated by $\{X_t : t \in T\}$. (1.3) can be equivalently defined as, for any $f \in \mathcal{B}_b(E)$ and all $s, t \in T$ with $s < t$,

$$\mathbb{E}[f(X_t) \mid \mathcal{F}_s] = \mathbb{E}[f(X_t) \mid X_s].$$

Example 1. *Standard Brownian motion.* An important example of a Markov process is a *standard Brownian motion* or *Wiener process.* $1-$dimensional Brownian motion $B_t : \mathbb{R}_+ \to \mathbb{R}$ is a real-valued stochastic process such that:

i) $B_0 = 0$ w.p.1.

ii) $t \mapsto B_t$ is continuous w.p.1.

iii) For any $0 \leq s < t$, $B_t - B_s$ is independent and distributed according to a Gaussian distribution with mean 0 and variance $t - s$.

Let $\mathcal{F}_n$ be the filtration generated by B_n. Then for any $f \in \mathcal{B}_b(\mathbb{R})$, we have $\mathbb{E}[f(B_t) \mid \mathcal{F}_s] = \mathbb{E}[f(B_t - B_s + B_s) \mid \mathcal{F}_s] = \mathbb{E}[g(B_t - B_s, B_s) \mid \mathcal{F}_s]$ where $g(x, y) = f(x + y)$. Since $B_t - B_s$ is is independent of $\mathcal{F}_s$ and B_s is $\mathcal{F}_s-$measurable, we conclude that $\mathbb{E}[g(B_t - B_s, B_s) \mid \mathcal{F}_s] = \mathbb{E}[g(B_t - B_s, B_s) \mid B_s] = \mathbb{E}[f(B_t) \mid B_s]$. Thus a standard Brownian motion is a Markov process.

Markov processes provide a theoretical basis not only for many modern computational methods but also for statistical modelling. Indeed, what we will see models and methods in the sequel of this paper are typically based on Markov processes. A transition kernel from $(E, \mathcal{E})$ into $(E, \mathcal{E})$ is called a *Markov kernel* on $(E, \mathcal{E})$ if $K(x, E) = 1$ for all $x \in E$. In other words, a Markov kernel is a probability transition kernel from $(E, \mathcal{E})$ into $(E, \mathcal{E})$. In this case, we can write $K : E \times \mathcal{E} \to [0, 1]$.

Let $\{P_{s,t}\}$ be a family of a Markov kernel on $(E, \mathcal{E})$. That is, assuming $X_s = x$, we have that $P_{s,t}(x, B) = \mathbb{P}(X_t \in B \mid X_s = x)$ for all $s, t \in T$ with $s \leq t$. Then a Markov process is said to be a *time-homogenous Markov process* if $\{P_{s,t}\}$ depends only on the difference $t-s$, i.e., if $\mathbb{P}(X_t \in B \mid X_s = x) = P_{s,t}(x, B) = P_{t-s}(x, B)$ holds. In the sequel, we will restrict our attention to time-homogenous case. In this case, the FDDs of a time-homogeneous Markov process starting at x at $t = 0$ are given by:

$$\begin{aligned}
\mathbb{P}(X_{t_1} \in B_1) &= P_{t_1}(x, B_1), \\
\mathbb{P}(X_{t_1} \in B_1, X_{t_2} \in B_2) &= \int_{B_1} P_{t_1}(x, dx_1) P_{t_2 - t_1}(x_1, B_2), \\
&\vdots \\
\mathbb{P}(X_{t_1} \in B_1, \cdots, X_{t_k} \in B_k) &= \int_{B_1} \cdots \int_{B_{k-1}} P_{t_1}(x, dx_1) \cdots P_{t_k - t_{k-1}}(x_{k-1}, dx_k).
\end{aligned}$$

This can be extended to a case when one has an initial distribution, say μ on E. In this case, we have that:

$$\mathbb{P}(X_{t_1} \in B_1, \cdots, X_{t_k} \in B_k) = \int_E \int_{B_1} \cdots \int_{B_{k-1}} \mu(dx) P_{t_1}(x, dx_1) \cdots P_{t_k - t_{k-1}}(x_{k-1}, dx_k).$$

Also, as a consequence of the Markovian property, we have the *Chapman-Kolmogorov equation.*

Proposition 2. *Let $\{X_t : t \in T\}$ be a time-homogenous Markov process on $(E, \mathcal{E})$. Then for any $s, t \in T$ with $s \leq t$ and $B \in \mathcal{E}$ we have that:*

$$P_t(x, B) = \int_E P_s(x, dy) P_{t-s}(y, B).$$

Proof. Due to the tower property of the conditional expectation, we have that:

$$\begin{aligned}
P_t(x, B) &= \mathbb{P}(X_t \in B \mid X_0 = x) = \mathbb{E}[\mathbb{I}\{X_t \in B\} \mid X_0 = x], \\
&= \mathbb{E}[\mathbb{E}[\mathbb{I}\{X_t \in B\} \mid X_0 = x] \mid \mathcal{F}_s] = \mathbb{E}[\mathbb{E}[\mathbb{I}\{X_t \in B\} \mid \mathcal{F}_s] \mid X_0 = x], \\
&= \mathbb{E}[\mathbb{P}(X_t \in B \mid X_s = y) \mid X_0 = x] = \mathbb{E}[P_{t-s}(y, B) \mid X_0 = x], \\
&= \int_E P_s(x, dy) P_{t-s}(y, B).
\end{aligned}$$

□

Therefore, as for a time-homogenous Markov process, the transition from time s to time t can be decomposed into two steps, that is, the first step goes from the initial state at time s to the intermediate state at time $t-s$. Then it moves from the intermediate state to the final state at time t.

The Chapman-Kolmogorov equation is one of the ubiquitous tools to analyse Markov processes. For instance, as we will study later, the filtering problem of hidden Markov models can be studied through the Chapman-Kolmogorov equation.

As we studied, we can define the operators for time-homogenous Markov processes as follows. Let $\mathcal{T}_t$ be the linear operator acting on $f \in \mathcal{B}_b(E)$ such that:

$$\mathcal{T}_t f(x) := \int_E f(y) P_t(x, dy) = \mathbb{E}\left[f(X_t) \mid X_0 = x\right], \tag{1.4}$$

for $t \in T$ with a notational convention $\mathcal{T}_0 = I$. From the Chapman-Kolmogorov equation and Fubini's theorem, it is clear to see that:

$$\begin{aligned}
\mathcal{T}_{t+s} f(x) &= \int_E f(y) P_{t+s}(x, dy) = \int_E \int_E f(y) P_s(z, dy) P_t(x, dz), \\
&= \int_E \left(\int_E f(y) P_s(z, dy) \right) P_t(x, dz) = \int_E \left(\mathcal{T}_s f(z)\right) P_t(x, dz), \\
&= \mathcal{T}_t \circ \mathcal{T}_s f(x),
\end{aligned}$$

and thus time-homogenous Markov processes can be studied via a semigroup of the operators $\{\mathcal{T}_t\}$, this is an example of the *Markov semigroup*. From Jensen's inequality, we have that:

$$\begin{aligned}
\left\|\mathcal{T}_t f(x)\right\|_\infty &= \left\| \int_E f(y) P_t(x, dy) \right\|_\infty \\
&\leq \left\|f(y)\right\|_\infty \int_E P(x, dy) = \left\|f(y)\right\|_\infty,
\end{aligned}$$

thus $\mathcal{T}_t$ is a contraction operator semigroup on $(\mathcal{B}_b(E), \|\|_\infty)$. Let $\mathcal{C}_b(E)$ be a space of continuous and bounded functions on E. For $f \in \mathcal{C}_b(E)$, denote by $\mathcal{D}(\mathcal{L})$ the set of all functions such that:

$$\mathcal{L} f := \lim_{t \downarrow 0} \frac{\mathcal{T}_t f - f}{t}, \tag{1.5}$$

exsist. The operator $\mathcal{L} : \mathcal{D}(\mathcal{L}) \to \mathcal{C}_b(E)$ is called the (infinitesimal) *generator* of the operator semigroup $\mathcal{T}_t$. The generator $\mathcal{L}$ is particularly useful for studying diffusion processes, as we will see later. Define $u(t, x) := \mathcal{T}_t f(x)$ and notice that one can write $\mathcal{T}_t = \exp\left(t\mathcal{L}\right)$. Then we have that:

$$\begin{aligned}
\nabla_t u(t, x) &= \nabla_t \mathcal{T}_t f(x) = \nabla_t \mathcal{T}_t \exp\left(t\mathcal{L} f\right), \\
&= \mathcal{L}\left(t\mathcal{L} f\right) = \mathcal{L}\mathcal{T}_t f(x) = \mathcal{L} u(t, x).
\end{aligned}$$

Since it is true that $u(0, x) = \mathcal{T}_0 f(x) = f(x)$, these give rise to:

$$\begin{cases}
\nabla_t u(t, x) = \mathcal{L} u(t, x), \\
u(0, x) = f(x).
\end{cases} \tag{1.6}$$

(1.6) is called the *backward Kolmogorov equation*. Clearly, the backward Kolmogorov equation determines the dynamics of the conditional expectation of (time-homogeneous) Markov processes. Also, we

can define the adjoint operator of $\mathcal{T}_t$, which acts for $\mu \in \mathcal{P}(E)$ and $B \in \mathcal{E}$ such that:

$$\mathcal{T}_t^* \mu(B) := \int_E \mu(dx) P(x, B) = \int_E \mu(dx) \mathbb{P}(X_t \in B \mid X_0 = x). \tag{1.7}$$

Indeed, we have that:

$$\int f(x) \mathcal{T}_t^* \mu(dx) = \int \mathcal{T}_t f(x) \mu(dx),$$

therefore $\mathcal{T}_t$ and $\mathcal{T}_t^*$ are adjoint operators. Also, as we studied, it can be shown that $\mathcal{T}_t^* \mu \in \mathcal{P}(E)$ for $\mu \in \mathcal{P}(E)$. Since $\mathcal{T}_t$ and $\mathcal{T}_t^*$ are adjoint, we can write $\mathcal{T}_t^* = \exp(t\mathcal{L}^*)$ where $\mathcal{L}^*$ is the adjoint operator of the generator $\mathcal{L}$. Suppose that a time-homogeneous Markov process has an initial distribution, say $\mu(dx) \in \mathcal{P}(E)$, in the sense that $X_0 \sim \mu(dx)$. We define:

$$\mu_t(dx) = \mathcal{T}_t^* \mu(dx). \tag{1.8}$$

This is the law of the (time-homogeneous) Markov process. Then, in the same manner as $\mathcal{L}$, we have $\nabla_t \mu_t = \mathcal{L}^* \mu_t$ and $\mu_0 = \mu$. Furthermore, assume that μ and μ_t admit the density denoted by p and p_t respectively w.r.t. the he Lebesgue measure. In this case, we have that:

$$\begin{cases} \nabla_t p_t = \mathcal{L}^* p_t, \\ p_0 = p. \end{cases} \tag{1.9}$$

This is the *forward Kolmogorov equation* or *the Fokker–Planck equation.* We note the if one sets $X_0 = x$ then $p_0 = \delta_x$.

1.2.3 Invariant distributions and ergodicity of Markov processes

As we noted, of particular interest is the long time behaviour of a time-homogeneous Markov process. In other words, we are interested in the stochastic stability of a time-homogeneous Markov process. Then an *invariant distribution* and *ergodicity* are two fundamental concepts to characterise such behaviour. We first introduce the definition of an invariant distribution.

Definition 7. Let $\{X_t : t \in T\}$ be a time-homogenous Markov process on $(E, \mathcal{E})$. Then $\mu(dx) \in \mathcal{P}(E)$ said to be an *invariant distribution* of the process if:

$$\mathcal{T}_t^* \mu(B) = \mu(B) \tag{1.10}$$

for any $B \in E$ and $t \in T$ holds.

In other words, an *invariant distribution is a fixed point* of $\mathcal{T}_t^*$. Assume that $\mu(dx)$ be an invariant

distribution and $X_0 \sim \mu(dx)$. Then FDDs of a time-homogenous Markov process are:

$$\begin{aligned}\mathbb{P}\left(X_{t_1+s} \in B_1, \cdots, X_{t_k+s} \in B_k\right) &= \int_E \int_{B_1} \cdots \int_{B_{k-1}} \mu(dx) P_{t_1+s}(x, dx_1) \cdots P_{t_k - t_{k-1}}(x_{k-1}, dx_k), \\ &= \int_{B_1} \cdots \int_{B_{k-1}} \mu(dx) P_{t_1}(x, dx_1) \cdots P_{t_k - t_{k-1}}(x_{k-1}, dx_k),\end{aligned}$$

therefore, the condition (1.10) implies that a time-homogeneous Markov process starting from $\mu(dx)$ has to be (strongly) stationary, and an invariant distribution is also called a stationary distribution of a Markov process to emphasise this fact. Also, above observation implies that *if* $X_0 \sim \mu(dx)$ then $X_t \sim \mu(dx)$ for any $t \in T$ with $t > 0$. Notice that dividing (1.10) by t and taking $t \downarrow 0$ give rise to:

$$\mathcal{L}^* p = 0. \tag{1.11}$$

This is often called the *stationary Fokker-Plank equation* in the context of diffusion processes.

Intuitively, an invariant distribution governs long time behaviour of a time-homogeneous Markov process. However, first of all, it is not clear to see whether an invariant distribution is a unique one. Also, an invariant distribution itself does not say anything about the situation in which $\{X_t : t \in T\}$ starts from an arbitrary initial distribution. That is, for $X_0 \sim \mu_0(dx)$, we are informally interested in whether:

$$\lim_{t \to \infty} \mathcal{T}_t^* \mu_0(dx) = \mu(dx),$$

holds in some sense, where $\mu(dx)$ is a unique invariant distribution of $\{X_t : t \in T\}$. This leads us to the following definition, called ergodicity.

Definition 8. (Ergodicity Pavliotis (2014, Chapter 2.4)). Let $\{X_t : t \in T\}$ be a time-homogenous Markov process on $(E, \mathcal{E})$. If there exists a unique distribution $\mu(dx)$ satisfying $\mathcal{T}_t^* \mu(dx) = \mu(dx)$, then the time-homogenous Markov process is said to be *ergodic* w.r.t. $\mu(dx)$.

Here we note that definition (8) can be obtained as a corollary of the definition of the ergodicity of a dynamical system in the context of a time-homogenous Markov process. See Hairer (Corollary 5.12, 2018) for instance. As a consequence of ergodicity, we have the following key theorem, as a consequence of Birkhoff's ergodic theorem.

Theorem 1. *Let $\{X_t : t \in T\}$ be a time-homogenous ergodic Markov process on $(E, \mathcal{E})$ with an arbitrary initial distribution and $f \in \mathcal{B}_b(E)$. Then we have that, w.p.1,*

$$\frac{1}{t} \int_0^t f(X_s) ds \to \int_E f(x) \mu(dx),$$

as $t \to \infty$, where $\mu(dx)$ is a unique invariant distribution.

Theorem 1 is also used as the definition of ergodicity, especially in physics (Pavliotis, 2014). Indeed, Definition 8 and Theorem 1 are equivalent. We note that, as we studied in subsubsection 1.2.1, if a time-homogenous Markov process is ergodic, then it will eventually become a process with no correlation

since the process will eventually be stationary. Thus the (auto) covariance of the process will decay as $t \to \infty$. Clearly, Theorem 1 says that the average w.r.t. time converges to the mathematical expectation on space, and indeed this gives us theoretical justification to use a (time-homogeneous) Markov process for simulation and modelling.

Although general theory for establishing uniqueness and existence of an invariant distribution of time-homogeneous Markov processes may be available, see Hairer (2010) for instance, the technicalities involved are beyond the scope of this thesis. Thus, to study uniqueness and existence of an invariant distribution, or ergodicity, we will restrict our attention to the two classes of (time-homogeneous) Markov processes in the following two successive subsections. That is, we will study the properties of *Markov chains* and *Diffusion processes* in what follows.

1.3 Markov chain

1.3.1 Basics of Markov chains

A Markov chain is a Markov process with a discrete time parameter $t \in \mathbb{Z}_+$, so we will take $T = \mathbb{Z}_+$ throughout this section. To be precise, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $\{X_t : t \in T\}$ be a stochastic process taking values on $(E, \mathcal{E})$. Let $\mathcal{F}_t^X$ be the filtration such that $\mathcal{F}_t := \sigma(X_t : j \leq t, j \in T)$. Then $\{X_t : t \in T\}$ is said to be a *Markov chain* if $\mathbb{P}(X_{t+1} \in B \mid \mathcal{F}_t) = \mathbb{P}(X_{t+1} \in B \mid X_t = x_t)$ holds (w.p.1) for any $B \in \mathcal{E}$. It can be shown that this definition is equivalent to the condition such that for any $f \in \mathcal{B}_b(E)$, $\mathbb{E}[f(X_{t+1}) \in B \mid \mathcal{F}_t] = \mathbb{E}[f(X_{t+1}) \in B \mid X_t = x]$ holds (w.p.1) for any $B \in \mathcal{E}$.

Definition 9. Let P be a Markov kernel on $(E, \mathcal{E})$. A Markov chain $\{X_t : t \in T\}$ is said to be a *time-homogeneous Markov chain* if $\mathbb{P}(X_{t+1} \in B \mid X_t = x) = P(x, B)$ holds for any $B \in \mathcal{E}$ and $t \in T$.

Applying Proposition 2 to a time-homogeneous Markov chain immediately gives rise to the following, we omit the proof.

Proposition 3. *Let $\{X_t : t \in T\}$ a time-homogenous Markov chain on $(E, \mathcal{E})$ with a Markov kernel P. Then we have that:*

$$\mathbb{P}(X_n \in B \mid X_0 = x) = P^n(x, B),$$

for any $B \in \mathcal{E}$, where

$$P^1 = P,\ P^n(x, B) = \int_E P(x, dy) P^{n-1}(y, B),$$

also

$$P^{n+k}(x, B) = \int_E P^n(x, dy) P^k(y, B),$$

holds for any $n, k \geq 1$.

Example 2. *Autoregressive model.* Consider the following example of a Markov chain:

$$X_t = \phi X_{t-1} + \epsilon_t,$$

where $\epsilon_t \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$ with $X_0 = x$. Then the Markov kernel (transition density) $p(x, x')$ is given by $\mathcal{N}(x'; \phi x, \sigma^2)$. Also we have $X_t = \phi(\phi X_{t-2} + \epsilon_{t-1}) + \epsilon_t = \cdots = \phi^t x + \sum_{n=0}^{t-1} \phi^n \epsilon_{t-n}$ so that $X_t \sim \mathcal{N}(\phi^t x, \sum_{n=0}^{t-1} \phi^{2n}\sigma^2)$ follows from the reproductive property of a Gaussian distribution. If $\mid \phi \mid < 1$, then we have $X_t \to \mathcal{N}\left(0, \frac{\sigma^2}{1-\phi^2}\right)$ as $t \to \infty$ in distribution. Also it can be shown that this chain is stationary.

Notice that we can write $P(x, B) = \mathbb{P}(X_{n+1} \in B \mid X_n = x)$ for any $n \in T$ for a time-homogenous Markov chain. As a consequence of Proposition 3, if one specifies an initial distribution v of X_0, then ν and P completely characterise the joint distribution of a time-homogenous Markov chain $\{X_t : t \in T\}$ in the sense that:

$$\mathbb{E}[f(X_{0:n})] = \int_{E^{n+1}} f(x_{0:n})\nu(dx_0)P(x_0, dx_1)\cdots P(x_{n-1}, dx_n),$$

for $f \in \mathcal{B}_b(E^{n+1})$. Also, the marginal distribution of X_n is given by:

$$\mu(dx_n) = \int \nu(dx_0)P^n(x_0, dx_n). \tag{1.12}$$

As before, we define the following operator on $\mathcal{P}(E)$.

Definition 10. Let P be a Markov kernel on $(E, \mathcal{E})$. For $\mu \in \mathcal{P}(E)$, we define the operator $\mathcal{T}^*$ on $\mathcal{P}(E)$ such that:

$$\mathcal{T}^*\mu(B) := \int_E \mu(dx)P(x, B), \tag{1.13}$$

for $B \in \mathcal{E}$.

The operator $\mathcal{T}^*$ is often called *left multiplication*. We note that if one takes $\mu(dx) = \delta_y(dx)$ then $\int_E \mu(dx)P(x, B) = P(y, B)$. Recall that $\mu \in \mathcal{P}(E)$ is said to be an *invariant distribution* (measure) if:

$$\mu(B) = \mathcal{T}^*\mu(B),$$

holds, in other words, $\int_E \mu(dx)P(x, B) = \mu(B)$ holds. In this case, we will also say a *Markov kernel* P *preserves* μ. Suppose that $\mu \in \mathcal{P}(E)$ is an invariant distribution and $X_0 \sim \mu(dx_0)$, then:

$$\begin{aligned}
\mu(B) &= \int_E \left[\int_E \mu(dx)P(x, dy)\right] P(y, B), \\
&= \int_E \mu(dx)P^2(x, B), \\
&\vdots \\
&= \int_E \mu(dx)P^n(x, B) = \mathbb{P}(X_n \in B),
\end{aligned}$$

holds for $B \in \mathcal{E}$ from Proposition 3 so that *a Markov chain* $\{X_t : t \in T\}$ *is strongly stationary if its initial distribution is an invariant one.* Equivalently $\mu(dx)$ is an invariant distribution of X_t if $X_0 \sim \mu(dx)$ then $X_1 \sim \mu(dx)$. We also introduce the following important definition.

Definition 11. Let P be a Markov kernel on $(E, \mathcal{E})$. We say P is *reversible* w.r.t. $\mu \in \mathcal{P}(E)$ if

$$\int_A \mu(dx)P(x, B) = \int_B \mu(dx)P(x, A), \tag{1.14}$$

holds for any $A, B \in \mathcal{E}$. Equivalently if:

$$\mu(dx)P(x, dy) = \mu(dy)P(y, dx),$$

holds.

Clearly, if P is reversible w.r.t. $\mu \in \mathcal{P}(E)$, then $\int_A \mu(dx)P(x, B) = \int_B \mu(dx)P(x, E) = \mu(B)$ so that P preserves μ. Thus *reversibility implies invariance.* Also we will say P *is* $\mu-$*reversible.* The condition (1.14) is also known as the *detailed balance condition.* Notice that if one takes $\mu \in \mathcal{P}(E)$ as an initial distribution of a a time-homogenous Markov chain $\{X_t : t \in T\}$, then the condition (1.14) implies:

$$\begin{aligned}\mathbb{E}\left[f(X_0, X_1)\right] &= \int f(x_0, x_1)\mu(dx_0)P(x_0, dx_1), \\ &= \int f(x_1, x_0)\mu(dx_0)P(x_0, dx_1),\end{aligned}$$

so X_0 and X_1 are *exchangeable* in the sense that the distribution of (X_0, X_1) is the same as of (X_1, X_0) hence the distribution of $(X_0, \cdots X_n)$ is the same as of $(X_n, \cdots X_0)$ for any $n \in T$ by induction.

Example 3. Again consider the autoregressive model (Example 2). Assume that $X_0 \sim \mathcal{N}\left(0, \frac{\sigma^2}{1-\phi^2}\right)$. Then $X_1 = \phi X_0 + \epsilon_1 \sim \mathcal{N}\left(0, \frac{\sigma^2}{1-\phi^2}\right)$ follows from the reproductive property of a Gaussian distribution. Thus $\mathcal{N}\left(0, \frac{\sigma^2}{1-\phi^2}\right)$ is an invariant distribution of the chain. Also the joint distribution of (X_0, X_1) is given by $\mathcal{N}\left(\begin{pmatrix} \mathbb{E}[X_0] \\ \mathbb{E}[X_1] \end{pmatrix}, \begin{pmatrix} \mathbb{V}[X_0] & \mathrm{Cov}(X_0, X_1) \\ \mathrm{Cov}(X_1, X_0) & \mathbb{V}[X_1] \end{pmatrix}\right)$. Since $\mathbb{E}[X_0] = \mathbb{E}[X_1]$ and $\mathbb{V}[X_0] = \mathbb{V}[X_1]$, the joint distribution of (X_0, X_1) is the same as the one of (X_1, X_0) so that the chain is $\mathcal{N}\left(0, \frac{\sigma^2}{1-\phi^2}\right)-$reversible and thus exchangeable.

In the context of reversible Markov chains, there are two important Hilbert spaces. Let $L^2(\pi)$ be the Hilbert space of real functions which are integrable with respect to a probability measure $\pi \in \mathcal{P}(E)$ such that:

$$\int_E f(x)^2\pi(dx) < +\infty \iff f \in L^2(\pi), \tag{1.15}$$

equipped with the inner product and associated norm:

$$\langle f, g \rangle := \int_E f(x)g(x)\pi(dx),$$
$$\|f\|_2 := \left(\int_E f(x)^2 \pi(dx) \right)^{1/2}.$$

We note that $L^2(\pi)$ should be understood as the equivalent class of the quotient space so that we will treat $\|f\|_2$ as a norm. We will not cover the completion issues. For $f \in L^2(\pi)$, we also define:

$$\begin{aligned} L_0^2(\pi) &:= \left\{ f \in L^2(\pi) : \int_E f(x)\pi(dx) = 0 \right\}, \\ &= \left\{ f \in L^2(\pi) : \langle f, 1 \rangle = 0 \right\}. \end{aligned} \tag{1.16}$$

Thus $L_0^2(\pi)$ is the subspace of $L^2(\pi)$ orthogonal to the constant functions. Since the linear function $f \mapsto \langle f, 1 \rangle$ is continuous, $L_0^2(\pi)$ is a closed subspace of $L^2(\pi)$ *and thus is also a Hilbert space*, see Kreyszig (1978, Theorem 3.2-4) for instance. Notice that for $f, g \in L_0^2(\pi)$, we have that $\mathrm{Cov}(f(X), g(X)) = \langle f, g \rangle$ and $\mathbb{V}(f(X)) = \langle f, f \rangle$. Then, as before, we define the *right multiplication* for $f \in \mathcal{B}_b(E)$:

$$\mathcal{T}f(x) := \int_E f(y)P(x, dy), \tag{1.17}$$

that is, $\mathcal{T}f(x) = \mathbb{E}\left[f(X_n) \mid X_{n-1} = x \right]$. We will also write $\mathcal{T}f$ depending on the context. First of all, one can observe that:

$$\begin{aligned} \int_E \mathcal{T}f(x)\mu(dx) &= \int_E \int_E f(y)P(x, dy)\mu(dx), \\ &= \int_E f(y) \int_E \mu(dx)P(x, dy) = \int_E f(x)\mathcal{T}^*\mu(dx), \end{aligned}$$

so that $\mathcal{T}$ and $\mathcal{T}^*$ are adjoint operators, as we mentioned. Then, the Jensen's inequality again gives:

$$\left[\mathcal{T}f(x) \right]^2 = \left[\int_E f(y)P(x, dy) \right]^2 \leq \int_E f(y)^2 P(x, dy).$$

Moreover, suppose that P is $\pi-$reversible and $f \in L^2(\pi)$, then we have that:

$$\begin{aligned} \|\mathcal{T}f(x)\|_2^2 &= \int_E \left[\mathcal{T}f(x) \right]^2 \pi(dx) \leq \int_E \int_E f(y)^2 P(x, dy)\pi(dx), \\ &= \int_E f(y)^2 \pi(dy) = \|f\|_2^2 < +\infty, \end{aligned}$$

holds by Fubini's theorem so that $\mathcal{T}$ *is a bounded linear (thus continuous) operator* in $L^2(\pi)$. Also,

critically, we can show that for $f, g \in L^2(\pi)$:

$$\begin{aligned}\langle \mathcal{T} f, g \rangle &= \int_E \int_E \pi(dx) P(x, dy) f(y) g(x) = \int_E \int_E \pi(dy) P(y, dx) f(y) g(x), \\ &= \int_E \int_E \pi(dy) P(y, dx) f(x) g(y) = \langle f, \mathcal{T} g \rangle ,\end{aligned}$$

holds if P is $\pi-$reversible so that $\mathcal{T} = \mathcal{T}^*$ in $L^2(\pi)$ meaning $\mathcal{T}$ *is a (Hilbert) self-adjoint operator* in $L^2(\pi)$. Thus we have the following.

Proposition 4. *A Markov kernel P is $\pi-$reversible iff $\mathcal{T}$ (1.17) is a self-adjoint operator in $L^2(\pi)$.*

1.3.2 Ergodicity of Markov chains

Recall that an invariant distribution is defined as a fixed point of the operator $\mathcal{T}^*$. Although there are several ways to establish ergodicity of Markov chains, one straightforward way to establish the uniqueness of an invariant distribution is using the Banach fixed-point theorem (also known as the contraction mapping theorem), see, e.g. Kreyszig (1978, Chapter 5). That is, we want to show that the operator $\mathcal{T}^* : \mathcal{P}(E) \to \mathcal{P}(E)$ has a unique fixed point. We refer to Hairer and Mattingly (2011); Douc et al. (2018); Hairer (2018) for the details of this approach.

To do so, clearly, some care is needed. To be precise, a normed space $(\mathcal{P}(E), \|\cdot\|)$ has to be a complete one with some norm $\|\cdot\|$, and $\mathcal{T}^*$ has to be contraction mapping on that space. As for the norm, we will use the *total variation norm*. Recall that any $\mu \in \mathcal{S}(E)$ can be decomposed (Hahn decomposition) as the difference of two positive measures (denoted by μ^+ and μ^-) with disjoint supports, that is they are singular. Then the *total variation norm* of $\mu \in \mathcal{S}(E)$ is given by:

$$\|\mu\|_{TV} := \mu^+(E) - \mu^-(E). \tag{1.18}$$

In the context of Markov chains, the following definition is routinely used as a definition of ergodicity of Markov chains.

Definition 12. Let $\{X_t : t \in T\}$ be a (time-homogenous) Markov chain on $(E, \mathcal{E})$ with a Markov kernel P and an invariant distribution $\pi \in \mathcal{P}(E)$. Then $\{X_t : t \in T\}$ is said to be *ergodic* if:

$$\lim_{n \to \infty} \|\mu P^n(\cdot) - \pi(\cdot)\|_{TV} = 0, \tag{1.19}$$

holds for *any initial distribution* $\mu \in \mathcal{P}(E)$.

We note that if a (time-homogenous) Markov chain is ergodic tine the sense, then Theorem 1 holds so that Definition 8 and Definition 12 are equivalent, see, e.g. Douc et al. (2018, Proposition 5.2.14). Also, it can be shown that $(\mathcal{S}(E), \|\|_{TV})$ is a normed vector space. Then the total variation norm induces the following distance.

Definition 13. Let $(E, \mathcal{E})$ be a measurable space. Then for any $\mu, \nu \in \mathcal{P}(E)$, the *total variation*

distance $d(\mu,\nu)_{TV}$ is given by:

$$d(\mu,\nu)_{TV} := \frac{1}{2}\left\|\mu-\nu\right\|_{TV} = \sup_{B\in\mathcal{E}}\left(\mu(B)-\nu(B)\right).$$

We note that the total variation distance is defined for $\mu,\nu\in\mathcal{S}(E)$. To see the meaning of the total variation distance, we assume that $\mu,\nu\in\mathcal{P}(E)$ admit densities (also denoted as μ,ν) w.r.t. the common Lebesgue measure, say dx so that $\frac{1}{2}\left\|\mu-\nu\right\|_{TV} = \sup_{B\in\mathcal{E}}\left|\int_B\left(\mu(x)-\nu(x)\right)dx\right|$. Then we have the following useful proposition.

Proposition 5. *Let $(E,\mathcal{E})$ be a measurable space and assume that $\mu,\nu\in\mathcal{P}(E)$ admit densities w.r.t. the common Lebesgue measure dx. Then we have that:*

$$d(\mu,\nu)_{TV} = \frac{1}{2}\int_E \mid \mu(x)-\nu(x) \mid dx.$$

Proof. Let $A := \{x\in E : \mu(x)\geq\nu(x)\}$. Then we have that:

$$\begin{aligned}\int_E \mid \mu(x)-\nu(x) \mid dx &= \int_A\left(\mu(x)-\nu(x)\right)dx + \int_{E\setminus A}\left(\mu(x)-\nu(x)\right)dx,\\ &\leq 2\sup_{B\in\mathcal{E}}\left|\int_B\left(\mu(x)-\nu(x)\right)dx\right|.\end{aligned}$$

Also, since $\int_E\left(\mu(x)-\nu(x)\right)dx = \mu(E)-\nu(E) = 0 = \nu(E)-\mu(E) = \int_E\left(\nu(x)-\mu(x)\right)dx$, we obtain:

$$\int_A\left(\mu(x)-\nu(x)\right)dx = \int_{E\setminus A}\left(\mu(x)-\nu(x)\right)dx.$$

Besides, for any $B\in\mathcal{E}$, we have that:

$$\begin{aligned}\left|\int_B\left(\mu(x)-\nu(x)\right)dx\right| &= \max\left\{\int_B\left(\mu(x)-\nu(x)\right)dx, \int_B\left(\nu(x)-\mu(x)\right)dx\right\},\\ &\leq \max\left\{\int_{B\cap A}\left(\mu(x)-\nu(x)\right)dx, \int_{B\cap(E\setminus A)}\left(\nu(x)-\mu(x)\right)dx\right\},\\ &\leq \max\left\{\int_A\left(\mu(x)-\nu(x)\right)dx, \int_{E\setminus A}\left(\mu(x)-\nu(x)\right)dx\right\},\\ &= \int_A\left(\mu(x)-\nu(x)\right)dx = \frac{1}{2}\int_E \mid \mu(x)-\nu(x) \mid dx,\end{aligned}$$

here we used $\int_E \mid \mu(x)-\nu(x) \mid dx = 2\int_A\left(\mu(x)-\nu(x)\right)dx$. Taking the supremum over $B\in\mathcal{E}$ gives rise to:

$$\sup_{B\in\mathcal{E}}\left|\int_B\left(\mu(x)-\nu(x)\right)dx\right| \leq \frac{1}{2}\int_E \mid \mu(x)-\nu(x) \mid dx,$$

thus $\frac{1}{2}\int_E \mid \mu(x)-\nu(x) \mid dx = \sup_{B\in\mathcal{E}}\left|\int_B\left(\mu(x)-\nu(x)\right)dx\right| = \frac{1}{2}\left\|\mu-\nu\right\|_{TV}$. □

Remark 1. *Recall that we say $\mu,\nu\in\mathcal{P}(E)$ are singular if there exists $A\in\mathcal{E}$ such that $\mu(A) = \nu(A^c) =$*

0, *and write* $\mu \perp \nu$ *in such case. Then it can be shown that* $\mu \perp \nu$ *is equivalent to* $d(\mu,\nu)_{TV} = 2$*. Also, it turns out that* $d(\mu,\nu)_{TV} < 2$ *is equivalent to* $\mu, \nu \in \mathcal{P}(E)$ *are not singular.*

Thus, it turns out that the total variation distance $d(\mu,\nu)_{TV}$ is just the L^1 distance between $\mu, \nu \in \mathcal{P}(E)$ when they admit densities. Upon observing this proposition, we immediately obtain the following lemma, we omit the proof since this follows from the completeness of L^1 space, which is well known, see Hairer (2018, Lemma 4.28) for instance.

Lemma 1. *Let* $(E, \mathcal{E})$ *be a mesurable space and assume that* $\mu, \nu \in \mathcal{P}(E)$ *admit densities w.r.t. the common Lebesgue measure* dx*. Then the space* $(\mathcal{P}(E), \|\|_{TV})$ *is Banach space and the space* $(\mathcal{P}(E), d_{TV})$ *is Polish space.*

Uniform ergodicity The definition of ergodicity in (1.19) does not say anything about the rate of this convergence. We first introduce the *Dobrushin coefficient.*

Definition 14. (Dobrushin coefficient). Let $\{X_t : t \in T\}$ be a (time-homogenous) Markov chain on $(E, \mathcal{E})$ with a Markov kernel P. Then the *Dobrushin coefficient* $\Delta(P)$ is given by:

$$\Delta(P) := \sup_{\mu \neq \nu \in \mathcal{P}(E)} \frac{d_{TV}(\mu P, \nu P)}{d_{TV}(\mu, \nu)} = \sup_{\mu \neq \nu \in \mathcal{P}(E)} \frac{\|\mu P - \nu P\|_{TV}}{\|\mu - \nu\|_{TV}}. \tag{1.20}$$

Notice that for two Markov kernels P, Q, it can be shown that:

$$d_{TV}(\mu PQ, \nu PQ) \leq \Delta(Q) d_{TV}(\mu P, \nu P) \leq \Delta(Q)\Delta(P) d_{TV}(\mu, \nu), \tag{1.21}$$

so that:

$$\Delta(PQ) \leq \Delta(P)\Delta(Q). \tag{1.22}$$

Moreover, taking $\mu, \nu = \delta_x, \delta_{x'}$ gives rise to the following (Douc et al., 2018, Lemma 18.2.2).

Lemma 2. *Let* $\{X_t : t \in T\}$ *be a (time-homogenous) Markov chain on* $(E, \mathcal{E})$ *with a Markov kernel* P*. Then:*

$$\Delta(P) = \sup_{x, x' \in E} d_{TV}(P(x, \cdot), P(x', \cdot)) \leq 1.$$

The following convergence rate property is called *uniform ergodicity.*

Definition 15. (Uniform ergodicity). Let $\{X_t : t \in T\}$ be a time-homogenous Markov chain on $(E, \mathcal{E})$ with a Markov kernel P and an invariant distribution $\pi \in \mathcal{P}(E)$. Then $\{X_t : t \in T\}$ is called *uniformly ergodic* if there exists constants $C < \infty$ and $\rho \in (0, 1)$ such that:

$$\sup_{x \in E} \|P^n(x, \cdot) - \pi(\cdot)\|_{TV} \leq C\rho^n, \tag{1.23}$$

holds for any $n \in T$.

Clearly, (1.23) implies that $\lim_{n\to\infty}\sup_{x\in E}\|P^n(x,\cdot)-\pi(\cdot)\|_{TV}=0$ since a constant $C<\infty$ does not depend on $x\in E$. Therefore, uniform ergodicity requires a Markov chain $\{X_t : t\in T\}$ uniformly geometrically converges to a invariant distribution $\pi\in\mathcal{P}(E)$ in terms of total variation distance. Indeed, it can be shown that $\lim_{n\to\infty}\sup_{x\in E}\|P^n(x,\cdot)-\pi(\cdot)\|_{TV}=0$ implies (1.23), see e.g. Douc et al. (2018, Chapter 18). The following lemma says *if $\Delta(P)<1$, then a (time-homogenous) Markov chain will forget its initial distribution (position) exponentially fast.*

Lemma 3. *Let $\{X_t : t\in T\}$ be a (time-homogenous) Markov chain on $(E,\mathcal{E})$ with a Markov kernel P. Then for any $\mu,\nu\in\mathcal{P}(E)$, the sequence $\{d_{TV}(\mu P^n,\nu P^n); n\in T\}$ is non-increasing and:*

$$d_{TV}(\mu P^n,\nu P^n)\le\Delta(P)^n d_{TV}(\mu,\nu). \tag{1.24}$$

If $\pi\in\mathcal{P}(E)$ is an invariant distribution, then for any $\mu\in\mathcal{P}(E)$, the sequence $\{d_{TV}(\mu P^n,\pi P^n); n\in T\}$ is decreasing and $d_{TV}(\mu P^n,\pi)\le\Delta(P)^n d_{TV}(\mu,\pi)$.

Proof. From Proposition 3 and (1.21), we have that $d_{TV}(\mu P^{n+1},\nu P^{n+1})\le\Delta(P)d_{TV}(\mu P^n,\nu P^n)$ holds since $\Delta(P)\le 1$ (Lemma 2), so that $\{d_{TV}(\mu P^n,\nu P^n); n\in T\}$ is non-increasing. Then (1.24) follows by induction. Taking $\nu=\pi$ gives rise to the rest of the claim since $\pi P^n=\pi$. □

Again, if $\Delta(P)<1$ then $\lim_{n\to\infty}d_{TV}(\mu P^n,\nu P^n)=0$ holds exponentially fast for any $\mu,\nu\in\mathcal{P}(E)$. This property of Markov chains is known as *the forgetting property*. From Lemma 3 and the Banach fixed-point theorem, we immediately have the following.

Theorem 2. *Let $\{X_t : t\in T\}$ be a (time-homogenous) Markov chain on $(E,\mathcal{E})$ with a Markov kernel P such that for some integer n, $\Delta(P^n)\le\rho<1$ holds. Then P admits a unique invariant distribution $\pi\in\mathcal{P}(E)$, and for any $\mu\in\mathcal{P}(E)$, $d_{TV}(\mu P^m,\pi)\le\rho^{\lfloor m/n\rfloor}d_{TV}(\mu,\pi)$ holds where $\lfloor x\rfloor$ is the floor function.*

Since the total variation distance of two probability measures is always less than 1, under the assumptions in Theorem 2, we have that:

$$d_{TV}(\mu P^m,\pi)\le(1-\rho)^{\lfloor m/n\rfloor}. \tag{1.25}$$

This implies that the convergence is uniform w.r.t. any initial distributions $\mu\in\mathcal{P}(E)$ if Theorem 2 holds. Therefore, we next introduce the sufficient condition, known as the *Doeblin/Minorisation condition*, of Theorem 2.

Definition 16. (Doeblin condition). Let $\{X_t : t\in T\}$ be a (time-homogenous) Markov chain on $(E,\mathcal{E})$ with a Markov kernel P. Then P satisfies the *Doeblin condition* if there exists an integer $n\ge 1$, $\epsilon>0$ and a probability measure $\nu\in\mathcal{P}(E)$ such that for any $x\in X$ and $B\in\mathcal{E}$:

$$P^n(x,B)\ge\epsilon\nu(B). \tag{1.26}$$

Remark 2. *If (1.26) holds, then such ϵ is necessarily $\epsilon\in(0,1]$. Indeed, taking $B=E$ gives rise to $1=P^n(x,E)\ge\epsilon\nu(E)=\epsilon$.*

Proposition 6. *Let $\{X_t : t \in T\}$ be a (time-homogenous) Markov chain on $(E, \mathcal{E})$ with a Markov kernel P. Assume that P satisfies the Doeblin condition (1.26). Then we have that $\Delta(P^n) \leq 1 - \epsilon$. That is, the chain is uniformly ergodic.*

Proof. Define $\tilde{P}(x, B) := (1 - \epsilon)^{-1} \{P(x, B) - \epsilon\nu(B)\}$ for any $x \in E$, $B \in \mathcal{E}$. The Doeblin condition guarantees that $\tilde{P}$ is a Markov kernel, and we have $\mu P - \mu' P = (1 - \epsilon) \left\{\mu\tilde{P} - \mu'\tilde{P}\right\}$. Therefore, in particular, we have:

$$\begin{aligned}\frac{1}{2} \left\|P^n(x, \cdot) - P^n(x', \cdot)\right\|_{TV} &= \frac{1}{2}(1 - \epsilon) \left\|\tilde{P}^n(x, \cdot) - \tilde{P}^n(x', \cdot)\right\|_{TV}, \\ &\leq (1 - \epsilon)\Delta(\tilde{P}^n) \leq 1 - \epsilon,\end{aligned}$$

and thus we have $\Delta(P^n) \leq 1 - \epsilon$ since the last inequality above does not depend on $x \in X$. □

Geometric ergodicity The notion of uniform ergodicity is restrictive in practice, especially when a state space of a chain is a non-compact set. Indeed, consider again the autoregressive model (Example 2) with $\epsilon_n \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$. Then $n-$step transition density $p^n(x, \cdot)$ is given by $\mathcal{N}\left(\phi^n x, \frac{1-\phi^{2n}}{1-\phi^2}\right)$, and an invariant density π is given by $\mathcal{N}\left(0, \frac{1}{1-\phi^2}\right)$. Therefore, from Proposition 5, we have:

$$\sup_{x \in \mathbb{R}} \left\|p^n(x, \cdot) - \pi\right\|_{TV} = 1.$$

Thus the convergence is not uniform. This motivate us to define the following convergence rate property.

Definition 17. (Geometric ergodicity). Let $\{X_t : t \in T\}$ be a time-homogenous Markov chain on $(E, \mathcal{E})$ with a Markov kernel P and an invariant distribution $\pi \in \mathcal{P}(E)$. Then $\{X_t : t \in T\}$ is called *geometrically ergodic* if there exists a constant $\rho \in (0, 1)$, and a non-negative function $M(x) : E \to [0, \infty]$ such that $\int_E M(x)\pi(dx) < \infty$ w.p.1. and for all $x \in E$ and $n \in T$:

$$\left\|P^n(x, \cdot) - \pi(\cdot)\right\|_{TV} \leq M(x)\rho^n.$$

Also, we introduce the following conditions.

Definition 18. (Geometric drift condition). Let $\{X_t : t \in T\}$ be a time-homogenous Markov chain on $(E, \mathcal{E})$ with a Markov kernel P. Then the kernel P satisfies a *geometric drift condition* (or a Foster-Lyapunov condition) if there are constants $0 < \lambda < 1$ and $b < \infty$, and a function $V : E \to [1, \infty)$, such that:

$$PV \leq \lambda V + b, \tag{1.27}$$

i.e. such that $\int_E P(x, dy)V(y) \leq \lambda V(x) + b$ for any $x \in E$.

Definition 19. (Local Dobrushin coefficient). Let $\{X_t : t \in T\}$ be a (time-homogenous) Markov chain on $(E, \mathcal{E})$ with a Markov kernel P. Then the *local Dobrushin coefficient* $\Delta^L(P, K)$ of the kernel

P an a set $K \subseteq E$ is is given by:

$$\begin{aligned}\Delta^L(P,K) &:= \sup_{x,y\in K} \|P(x,\cdot) - P(y,\cdot)\|_{TV} = \sup_{x\neq y\in K} \frac{\|\delta_x P - \delta_y P\|_{TV}}{\|\delta_x - \delta_y\|_{TV}} \\ &= \sup_{\mu\neq\nu\in\mathcal{P}(K)} \frac{\|\mu P - \nu P\|_{TV}}{\|\mu-\nu\|_{TV}}.\end{aligned} \tag{1.28}$$

The following result (Hairer and Mattingly, 2011; Eberle, 2020) shows that geometric ergodicity is a consequence of the Geometric drift condition and the local Doeblin condition.

Theorem 3. *(Eberle, 2020, Theorem 3.22),(Hairer and Mattingly, 2011, Theorem 3.1, 3.2). Assume that there exists a function $V : E \to [1,\infty)$ such that the geometric drift condition holds with constants λ, b and*

$$\Delta^L(P, \{x \in E; V(x) \leq r\}) < 1, \tag{1.29}$$

for some $r > 2b/(1-\lambda)$. Then there is a constant $\beta \in \mathbb{R}_+$ such that unique invariant probability measure π satisfying $\int_E V(x)\pi(dx) < \infty$, and geometric ergodicity holds:

$$\|P^n(x,\cdot) - \pi(\cdot)\|_{TV} \leq \mathcal{W}_\beta\left(P^n(x,\cdot), \pi(\cdot)\right) \leq \alpha_\beta^n \left(1 + \beta V(x) + \beta \int_E V(x)\pi(dx)\right),$$

where $\mathcal{W}_\beta(\mu,\nu)$ for $\mu,\nu \in \mathcal{P}(E)$ is defined as:

$$\begin{cases} \mathcal{W}_\beta(\mu,\nu) & := \inf_{X\sim\mu, Y\sim\nu} \mathbb{E}\left[d_\beta(\mu,\nu)\right], \\ d_\beta(x,y) & := 1_{\{x\neq y\}}\left(1 + \beta V(x) + \beta V(y)\right). \end{cases} \tag{1.30}$$

Proof. Given $x \neq y \in E$, selet (X,Y) such that $\mathbb{P}(X \neq Y) = \|\delta_x P - \delta_y P\|_{TV}$. Take $\lambda = 1-\gamma < 1$ for $\gamma > 0$ so that $r > 2b/\gamma$. Then for any $\beta > 0$, we have that:

$$\begin{aligned}\mathcal{W}_\beta(P(x,\cdot), P(y,\cdot)) &\leq \mathbb{P}(X \neq Y) + \beta\mathbb{E}\left[V(X)\right] + \beta\mathbb{E}\left[V(Y)\right], \\ &= \|\delta_x P - \delta_y P\|_{TV} + \beta PV(x) + \beta PV(y), \\ &\leq 1 + 2\beta b + \beta(1-\gamma)(V(x) + V(y)),\end{aligned}$$

from the geometric drift condition (1.27) and the fact that the total variation distance of two probability measures is always less than 1. Then we obtain:

$$\mathcal{W}_\beta(P(x,\cdot), P(y,\cdot)) \leq d_\beta(x,y) + 2\beta b - \beta\gamma(V(x) + V(y)),$$

assume that first $r \geq V(x) + V(y)$. Define $\delta := \frac{\beta(r\gamma - 2b)}{1+\beta r}$ and this is positive since $r\gamma - 2b > 0$. Then it can be shown that $\mathcal{W}_\beta(P(x,\cdot), P(y,\cdot)) \leq (1-\delta)d_\beta(x,y)$. Next assume that $r < V(x) + V(y)$. The from (1.29), one can show that, with the choices $\epsilon := \min\left\{\frac{1-\Delta^L(P,\{x\in E; V(x)\leq r\})}{2}, \lambda\right\}$ and $\beta \leq$

$\frac{1-\Delta^L(P,\{x\in E;V(x)\leq r\})}{4b}$:

$$\begin{aligned}\mathcal{W}_\beta(P(x,\cdot),P(y,\cdot)) &\leq \Delta^L(P,\{x\in E;V(x)\leq r\}) + 2b\beta + \beta(1-\gamma)(V(x)+V(y)),\\ &\leq (1-\epsilon)d_\beta(x,y).\end{aligned}$$

As a result, we have that:

$$\mathcal{W}_\beta(P(x,\cdot),P(y,\cdot)) \leq (1-\min\{\delta,\epsilon\})d_\beta(x,y),$$

for any $x\neq y\in E$. Thus there exists a unique stationary distribution π satisfying $\int_E V(x)\pi(dx)<\infty$ due to the Banach fixed-point theorem. Define $\alpha_\beta := (1-\min\{\delta,\epsilon\})<1$. Then we have that:

$$\begin{aligned}\mathcal{W}_\beta(P(x,\cdot)^n,\pi) = \mathcal{W}_\beta(\delta_x P^n,\pi P^n) &\leq \alpha_\beta^n \mathcal{W}_\beta(\delta_x,\pi),\\ &= \alpha_\beta^n\left(1+\beta V(x)+\beta\int_E V(x)\pi(dx)\right).\end{aligned}$$

The result follows from the fact that $\|\mu-\nu\|_{TV}\leq\mathcal{W}_\beta(\mu,\nu)$ with equality for $\beta=0$ for any $\mu\neq\nu\in\mathcal{P}(E)$. □

Consider again the autoregressive model (Example 2) with $\epsilon_n \overset{i.i.d.}{\sim} \mathcal{N}(0,\sigma^2)$. Define $V(x)=\mid x\mid+1$. Then $PV(x)\leq 1+\mid\phi\mid V(x)+\mathbb{E}[\mid\epsilon\mid]=1+\mid\phi\mid V(x)$. Thus (1.27) holds under the condition $\mid\phi\mid<1$. For large enough r, we have:

$$\Delta^L(P,\{x\in E;\mid x\mid+1\leq r\}) = \sup_{|x|\leq r-1}\sup_{|y|\leq r-1}\left\|\mathcal{N}(\phi x,\sigma^2)-\mathcal{N}(\phi y,\sigma^2)\right\|_{TV}\leq 1.$$

Then the autoregressive model is geometrically ergodic.

1.4 Diffusion process

1.4.1 Ito integrals

In this section, we focus on a special class of Markov process. A *diffusion process* is a (continuous time) Markov process with no jumps. In particular, we focus on Ito diffusions, following Øksendal (2003); Pavliotis (2014). To do so, we first need to consider the following integral, called the *Ito integral*:

$$\mathcal{I}(f) := \int_S^T f(t,\omega)dW_t(\omega), \tag{1.31}$$

where W_t is $1-$dimensional Brownian motion. We define the class of functions which Ito integrals are well defined.

Definition 20. (Øksendal, 2003, Definition 3.1.4). Let $\mathcal{V}$ be a class of functions $f(t,\omega):\mathbb{R}_+\times\Omega\to\mathbb{R}$ such that

i) $(t,\omega)\to f(t,\omega)$ is $\mathcal{B}(\mathbb{R}_+)\times\mathcal{F}-$measurable.

ii) Let $\mathcal{F}_t$ be the filtration generated by W_t. Then $f(t,\omega)$ is $\mathcal{F}_t$ adopted.

iii) $\mathbb{E}\left[\int_S^T f(t,\omega)^2 dW_t(\omega)\right] < \infty$.

For $f \in \mathcal{V}$, $\mathcal{I}(f)$ can be defined as the following 3 steps. First, let $g \in \mathcal{V}$ be a bounded and continuous function w.r.t. ω , and divide the interval $[S,T]$ as $t_0 = S < t_1 < \cdots t_n = T$. Then one can find a step function $\phi_n(t,\omega) := \sum_{j=1}^n g_j(t_j,\omega)1_{\{[t_j,t_{j+1})\}}(t)$ such that $\mathbb{E}\left[\int_S^T (g-\phi_n)^2 dt\right] \to 0$ as $n \to \infty$, where $1_{\{x\}}$ is an indicator function. Then let $h \in \mathcal{V}$ be bounded. Using the bounded convergence theorem, one can show that there exists $g_n \in \mathcal{V}$ such that $g(\cdot,\omega)$ is continuous for any n and ω, and $\mathbb{E}\left[\int_S^T (g_n - h)^2 dt\right] \to 0$ holds as $n \to \infty$. As a result, the dominated convergence theorem implies that there exists a sequence $\{h_n\} \in \mathcal{V}$ such that h_n is bounded for each n and $\mathbb{E}\left[\int_S^T (f-h_n)^2 dt\right] \to 0$ holds as $n \to \infty$ for $f \in \mathcal{V}$. Critically, it can be shown that $\mathbb{E}\left[\left(\int_S^T \phi_n(t,\omega)dW_t(\omega)\right)^2\right] = \mathbb{E}\left[\int_S^T \phi_n(t,\omega)^2 dt\right]$, and this implies that the sequence $\{\int_S^T \phi_n(t,\omega)dW_t(\omega)\}$ forms a Cauchy sequence in $L^2(\mathbb{P})$. Therefore, we finally can define the Ito integral.

Definition 21. (Ito integrals). Let $f \in \mathcal{V}$. Then *the Ito integral* of f is defined as $L^2(\mathbb{P})-$limit:

$$\mathcal{I}(f) = \lim_{n\to\infty} \int_S^T \phi_n(t,\omega)dW_t(\omega),$$

where $\{\phi_n\}$ is a sequence of step functions such that:

$$\mathbb{E}\left[\int_S^T \left(f(t,\omega) - \phi_n(t,\omega)\right)^2 dt\right] \to 0,$$

as $n \to \infty$.

Notice that the definition of the Ito integral is similar to the one of the Lebesgue integral. However, whilst the Lebesgue integral is defined as the almost sure limit, the Ito integral is defined as the $L^2(\mathbb{P})$ limit of the Cauchy sequence. Also, the multidimensional Ito integral can be defined, see Øksendal (2003, Chapter 3). We obtain the following important result.

Proposition 7. *(The Ito isometry). For any $f \in \mathcal{V}$, we have that:*

$$\mathbb{E}\left[\left(\int_S^T f(t,\omega)dW_t(\omega)\right)^2\right] = \mathbb{E}\left[\int_S^T f(t,\omega)^2 dt\right].$$

Proof. See Øksendal (2003, Corollary 3.1.7). □

Also we introduce a *martingale.*

Definition 22. (Martingale). Let $\{X_t : t \in T\}$ be a stochastic process and $\mathcal{F}_t$ be the filtration generated by $\{X_t : t \in T\}$. Then $\{X_t : t \in T\}$ is said to be a *martingale* w.r.t. $\mathcal{F}_t$ if

i) X_t is adopted w.r.t. $\mathcal{F}_t$.

ii) $\mathbb{E}\left[\mid X_t \mid\right] < \infty$ for all t.

iii) $\mathbb{E}\left[X_t \mid \mathcal{F}_s\right] = X_s$ for all $s \leq t$.

Clearly a Brownian motion W_t is a martingale w.r.t. the filtration $\mathcal{F}_t$ generated by $\{W_t : t \in T\}$. Indeed, $\mathbb{E}[W_t \mid \mathcal{F}_s] = \mathbb{E}[W_t - W_s + W_s \mid \mathcal{F}_s] = \mathbb{E}[W_t - W_s \mid \mathcal{F}_s] + W_s = W_s$.

Proposition 8. *For any $f \in \mathcal{V}$, $\int_0^t f(t,\omega)dW_s(\omega)$ is a martingale w.r.t. the filtration $\mathcal{F}_s$ generated by $\{W_s : s \in [0,t]\}$.*

Proof. See Øksendal (2003, Collorary 3.2.6). □

1.4.2 Basics of Ito diffusions

Definition 23. (Ito diffusions). A time homogeneous *Ito diffusion* is a stochastic process $X_t(\omega) : [0,T] \times \Omega \to \mathbb{R}^d$ satisfying a *stochastic differential equation* (SDE) of the form:

$$dX_t = b(X_t)dt + \sigma(X_t)dW_t, \tag{1.32}$$

with $X_0 = x$, where W_t is a $m-$dimensional Brownian motion and measurable functions $b : \mathbb{R}^d \to \mathbb{R}^d$, $\sigma : \mathbb{R}^d \to \mathbb{R}^{d\times m}$ satisfy:

$$\mid b(x) - b(y) \mid + \mid \sigma(x) - \sigma(y) \mid \leq C \mid x - y \mid, \tag{1.33}$$

for any $x, y \in \mathbb{R}^d$ and some constant $C > 0$, that is b and σ are Lipschitz continuous.

Notice that from (1.33) we have $\mid b(x) \mid \leq \mid b(0) \mid + C \mid x \mid \leq \{\mid b(0) \mid + C\}(1 + \mid x \mid)$ and $\mid \sigma(x) \mid \leq \mid \sigma(0) \mid + C \mid x \mid \leq \{\mid \sigma(0) \mid + C\}(1 + \mid x \mid)$. The following theorem guarantees existence and uniqueness of the solution to (1.32).

Theorem 4. *Assume that (1.33) holds. Then there exists an unique X_t such that X_t is adopted w.r.t. the filtration $\mathcal{F}_t$ generated by $\{W_t : t \in [0,T]\}$ and:*

i) *$X_0 = x$ w.p.1.*

ii) *$\int_0^t \{\mid b(X_s) \mid + \mid \sigma(X_s) \mid^2\} ds < \infty$ w.p.1.*

iii) *(1.32) holds w.p.1, that is $X_t = x + \int_0^t b(X_s)ds + \int_0^t \sigma(X_s)dW_s$ holds w.p.1.*

Proof. See Øksendal (2003, Theorem 5.2.1). □

The meaning of uniqueness in Theorem 4 is that if X_t and Y_t are strong solutions to (1.32), then $X_t = Y_t$ for any t w.p.1. The solution of the SDE (1.32) has the Markovian property and continuous path so that it is indeed a diffusion process, see Øksendal (2003, Chapter 8) for details. In the case of Ito diffusions, the generator $\mathcal{L}$ in (1.5) is defined as:

$$\mathcal{L} = \sum_{i=1}^{d} b_i(x)\frac{\partial}{\partial x_i} + \frac{1}{2}\sum_{i,j=1}^{d} \Sigma_{i,j}(x)\frac{\partial^2}{\partial x_i \partial x_j}, \tag{1.34}$$

where $\Sigma(x) := \sigma(x)\sigma(x)^\top$. In many scenarios, one might want to know the expression of a new process $V_t = g(t, X_t)$ where $g : [0,T] \times \mathbb{R}^d \to \mathbb{R}^d$ is a twice continuously differentiable function and X_t is an Ito diffusion. Using the operator in (1.34), the following theorem shows that how to derive the expression of dV_t.

Theorem 5. ***(Ito's lemma).*** *Let X_t be an Ito diffusion, and $g : [0,T] \times \mathbb{R}^d \to \mathbb{R}^d$ be a twice continuously differentiable function. Then the process $V_t = g(t, X_t)$ is also an Ito diffusion satisfying:*

$$V(t, X_t) = V(X_0) + \int_0^T \frac{\partial V}{\partial s}(s, X_s)ds + \int_0^T \mathcal{L}V(s, X_s)ds + \int_0^T \langle \nabla V(s, X_s), \sigma(X_s)dW_s \rangle ,$$

or equivalently:

$$dV(t, X_t) = \frac{\partial V_t}{\partial t}dt + \sum_{i=1}^{d} \frac{\partial V_t}{\partial x_i}dX_i + \frac{1}{2}\sum_{i,j=1}^{d} \frac{\partial^2 V}{\partial x_i \partial x_j}dX_i dX_j, \tag{1.35}$$

where $dW_i dW_j = \delta_{ij}dt$,$dW_i dt = dt dW_i = dt dt = 0$.

Proof. We refer to Øksendal (2003, Chapter 4) for a rigorous proof. Here we provide an informal proof when $d = 1$. First divide the interval $[0,T]$ as $t_0 = 0 < t_1 < \cdots t_n = T$. A Taylor series expansion of $V_t = g(t, X_t)$ gives rise to:

$$V_T = V_0 + \lim_{n\to\infty} \left[\sum_{i=1}^{n} \frac{\partial g}{\partial t}(t_i - t_{i-1}) + \sum_{i=1}^{n} \frac{\partial g}{\partial x}(X_{t_i} - X_{t_{i-1}}) + \frac{1}{2}\sum_{i=1}^{n} \frac{\partial^2 g}{\partial t^2}(t_i - t_{i-1})^2 + \sum_{i=1}^{n} \frac{\partial^2 g}{\partial t \partial x}(t_i - t_{i-1})(X_{t_i} - X_{t_{i-1}}) + \frac{1}{2}\sum_{i=1}^{n} \frac{\partial^2 g}{\partial x^2}(X_{t_i} - X_{t_{i-1}})^2 + \cdots \right],$$

where g are evaluated at (t_i, X_{t_i}). Also, it might be shown that:

$$\begin{aligned}\lim_{n\to\infty} \left[\sum_{i=1}^{n} \frac{\partial g}{\partial x}(X_{t_i} - X_{t_{i-1}}) \right] &= \lim_{n\to\infty} \left[\sum_{i=1}^{n} \frac{\partial g}{\partial x} b(X_{t_i})(t_i - t_{i-1}) + \sum_{i=1}^{n} \frac{\partial g}{\partial x}\sigma(X_{t_i})(W_{t_i} - W_{t_{i-1}}) \right], \\ &= \int_0^T \frac{\partial g}{\partial x} b(X_s)ds + \int_0^T \frac{\partial g}{\partial x}\sigma(X_s)dW_s,\end{aligned}$$

and:

$$\begin{aligned}\lim_{n\to\infty} \left[\sum_{i=1}^{n} \frac{\partial^2 g}{\partial x^2}(X_i - X_{i-1})^2 \right] &= \lim_{n\to\infty} \left[\sum_{i=1}^{n} \frac{\partial^2 g}{\partial x^2} b(X_{t_i})^2 (t_i - t_{i-1})^2 + \sum_{i=1}^{n} \frac{\partial^2 g}{\partial x^2}\sigma(X_{t_i})^2 (W_{t_i} - W_{t_{i-1}})^2 \right. \\ &\quad \left. + 2\sum_{i=1}^{n} \frac{\partial^2 g}{\partial x^2} b(X_{t_i})\sigma(X_{t_i})(t_i - t_{i-1})(W_{t_i} - W_{t_{i-1}}) \right], \\ &= \int_0^T \frac{\partial^2 g}{\partial x^2}\sigma(X_s)dt,\end{aligned}$$

since $(W_{t_i} - W_{t_{i-1}})^2 \to dt$ as $t_i - t_{i-1} \to 0$, and $(t_i - t_{i-1})^2 \to 0$ and $(t_i - t_{i-1})(W_{t_i} - W_{t_{i-1}}) \to 0$ as $t_i - t_{i-1} \to 0$. As a result, we have that:

$$V_T = V_0 + \int_0^T \left(\frac{\partial g(t, X_t)}{\partial t} + \frac{\partial g(t, X_t)}{\partial x} b(X_t) + \frac{1}{2}\frac{\partial^2 g(t, X_t)}{\partial x^2}\sigma^2(X_t) \right) ds + \int_0^T \frac{\partial g}{\partial x}\sigma(X_t)dW_t.$$

□

Example 4. *Integration by parts.* Let X_t and Y_t be Ito diffusions. Consider $g(x,y) = xy$. Then applying Theorem 5 to this function gives rise to

$$\begin{aligned} d(X_tY_t) &= \frac{\partial g(x,y)}{\partial x}dX_t + \frac{\partial g(x,y)}{\partial y}dY_t + \frac{1}{2}\frac{\partial^2 g(x,y)}{\partial x^2}(dX_t)^2 \\ &+ \frac{\partial^2 g(x,y)}{\partial x \partial y}dX_t dY_t + \frac{1}{2}\frac{\partial^2 g(x,y)}{\partial y^2}(dY_t)^2, \\ &= Y_t dX_t + X_t dY_t + dX_t dY_t. \end{aligned}$$

Thus we have that:

$$X_tY_t = X_0Y_0 + \int_0^t Y_s dX_s + \int_0^t X_s dY_s + \int_0^t dX_s dY_s, \tag{1.36}$$

and this is called the *integration by parts formula.*

Example 5. *Ornstein–Uhlenbeck process.* Consider the following scalar SDE:

$$dX_t = (\theta - \mu X_t)dt + \sigma dW_t,$$

with $X_0 = x$. Then the generator $\mathcal{L}$ of the process if given by:

$$\mathcal{L} = (\theta - \mu x)\frac{d}{dx} + \frac{\sigma^2}{2}\frac{d^2}{dx^2}.$$

Take $g(t,x) = x\exp(\mu t)$. Then Theorem 5 implies:

$$\begin{aligned} dX_t \exp(\mu t) &= \frac{d}{dt}X_t \exp(\mu t)\, dt + \mathcal{L}\left(X_t \exp(\mu t)\right) dt + \nabla\left(X_t \exp(\mu t)\right)\sigma dW_t, \\ &= \mu X_t \exp(\mu t)\, dt + (\theta - \mu X_t)\exp(\mu t)\, dt + \exp(\mu t)\,\sigma dW_t, \\ &= \theta \exp(\mu t)\, dt + \exp(\mu t)\,\sigma dW_t, \end{aligned}$$

and thus:

$$\begin{aligned} X_t \exp(\mu t) &= x + \theta\left(\frac{\exp(\mu t)}{\mu} - \frac{1}{\mu}\right) + \sigma\int_0^t \exp(\mu s)\, dW_s, \\ \longleftrightarrow X_t &= \frac{\theta}{\mu} + \left(x - \frac{\theta}{\mu}\right)\exp(-\mu t) + \sigma\int_0^t \exp\left(-\mu(t-s)\right) dW_s. \end{aligned}$$

Example 6. *Geometric Brownian motion.* Consider the following scalar SDE:

$$dX_t = \mu X_t dt + \sigma X_t dW_t,$$

with $X_0 = x$. Then the generator $\mathcal{L}$ is given by:

$$\mathcal{L} = \mu x\frac{d}{dx} + \frac{\sigma^2 x^2}{2}\frac{d^2}{dx^2}.$$

Take $g(t,x) = \log(x)$. Then Theorem 5 implies:

$$\begin{aligned} d\log X_t &= \mathcal{L}(\log X_t)\,dt + \nabla(\log X_t)\,\sigma X_t dW_t, \\ &= \left(\mu X_t \frac{1}{X_t} + \frac{\sigma^2 X_t^2}{2}\left(-\frac{1}{X_t^2}\right)\right)dt + \frac{1}{X_t}\sigma X_t dW_t, \\ &= \left(\mu - \frac{\sigma^2}{2}\right)dt + \sigma dW_t, \end{aligned}$$

and thus:

$$\begin{aligned} \log\frac{X_t}{X_0} &= \left(\mu - \frac{\sigma^2}{2}\right)t + \sigma W_t, \\ \longleftrightarrow X_t &= x\exp\left(\left(\mu - \frac{\sigma^2}{2}\right)t + \sigma W_t\right). \end{aligned}$$

As a particular and important application of Ito's lemma from a statistical point of view, one might want to remove $\sigma(X_t)$ from an Ito diffusion. Suppose that now one has a $1-$dimensional Ito diffusion X_t. To obtain an SDE with unit diffusion coefficient $\sigma(\cdot) = 1$, consider a transform $y = h(x)$. From Ito's lemma, we have that $dY_t = \mathcal{L}h(X_t)dt + \nabla h(X_t)\sigma(X_t)dW_t$. This implies the condition $\nabla h(X_t)\sigma(X_t) = 1$ to obtain the required SDE, and we have:

$$h(x) = \int_z^x \frac{1}{\sigma(x)}dx, \tag{1.37}$$

where z is any arbitrary value of X_t. Applying Ito's lemma to the map $h(x)$ yields $\mathcal{L}h(x) = \frac{b(x)}{\sigma(x)} - \frac{\nabla\sigma(x)}{2}$ so that we have:

$$\begin{cases} dY_t &= b_Y(Y_t)dt + dW_t, \\ b_Y(y) &:= \frac{b(h^{-1}(y))}{\sigma(h^{-1}(y))} - \frac{\nabla\sigma(h^{-1}(y))}{2}. \end{cases} \tag{1.38}$$

We summarise the transformation known as the *Lamperti transform*.

Proposition 9. ***(Lamperti transform).*** *Let X_t be $1-$dimensional Ito diffusion. Define the map $h(x) = y$ as in (1.37) Then the process solving the SDE in (1.38) has the same law as X_t.*

Example 7. *Cox-Ingersoll-Ross process.* Consider the following process:

$$dX_t = (\mu - \alpha X_t)dt + \sigma\sqrt{X_t}dW_t,$$

with $X_0 = x$. Then the transform $h(x)$ in (1.37) is given by $h(x) = \frac{2}{\sigma}\sqrt{x}$. The generator $\mathcal{L}$ of the process is given by $\mathcal{L} = (\mu - \alpha x)\frac{d}{dx} + \frac{\sigma^2 x}{2}\frac{d^2}{dx^2}$. As a result, we obtain:

$$\mathcal{L}h(x) = \left(\frac{\mu}{\sigma} - \frac{\sigma}{4}\right)x^{-1/2} - \frac{\alpha}{\sigma}x^{1/2}.$$

Then for $Y_t = \frac{2}{\sigma}\sqrt{X_t}$, we have:

$$\begin{aligned} dY_t &= \left(\frac{\mu}{\sigma} - \frac{\sigma}{4}\right) X_t^{-1/2} dt - \frac{\alpha}{\sigma} X_t^{1/2} dt + dW_t, \\ &= \left(\frac{2\mu}{\sigma^2} - \frac{1}{2}\right) \frac{1}{Y_t} dt - \frac{\alpha}{2} Y_t dt + dW_t. \end{aligned}$$

In practice, it is important to know the expression of a likelihood function, which is defined as the Radon–Nikodym derivative between two probability measures. The following theorem, known as *the Girsanov theorem*, describes the Radon–Nikodym derivative between two probability measures induced by two different Ito diffusions. Roughly speaking, the Girsanov theorem says that if one changes $b(\cdot)$ in an Ito diffusion, then the law of the process will not change dramatically. For the sake of simplicity, here we only consider the case when $d = 1$.

Theorem 6. ***(The Girsanov theorem).*** *Let $\mathbb{P}^{(1)}$ be the law induced by the Ito diffusion solving $dX_t = b^{(1)}(X_t)dt + \sigma(X_t)dW_t^{(1)}$, and $\mathbb{P}^{(2)}$ be the one induced by the Ito diffusion solving $dX_t = b^{(2)}(X_t)dt + \sigma(X_t)dW_t^{(2)}$ with $X_0 = x$. Define:*

$$u(x) := \frac{b^{(2)}(x) - b^{(1)}(x)}{\sigma(x)}, \tag{1.39}$$

and assume that:

$$\mathbb{E}_{\mathbb{P}^{(1)}}\left[\exp\left(\frac{1}{2}\int_0^t u(X_s)^2\right) ds\right] < \infty. \tag{1.40}$$

Also define:

$$\begin{aligned} \mathcal{G}(X_t) &:= \exp\left(\int_0^t u(X_s) dW_s^{(1)} - \frac{1}{2}\int_0^t u(X_s)^2 ds\right), \\ &= \exp\left(\int_0^t \frac{b^{(2)}(X_s) - b^{(1)}(X_s)}{\sigma(X_s)} dX_s + \int_0^t \frac{b^{(1)}(X_s)^2 - b^{(2)}(X_s)^2}{2\sigma(X_s)^2} ds\right). \end{aligned} \tag{1.41}$$

Then $\mathbb{P}^{(1)}$ and $\mathbb{P}^{(2)}$ are equivalent measures with the Radon-Nikodym derivative given by:

$$\frac{d\mathbb{P}^{(2)}}{d\mathbb{P}^{(1)}}\left(X_{[0,t]}\right) = \mathcal{G}(X_t). \tag{1.42}$$

Proof. We refer to Øksendal (2003, Chapter 8) for a rigorous proof. Here we provide an informal proof of the expression in (1.42). First divide the interval $[0, T]$ into M intervals with size $\delta = \frac{T}{M}$. Due to the Markovian property, the discretised joint density of $(X_{1\delta}, X_{2\delta}, \cdots X_{M\delta})$ under $\mathbb{P}_1$, denoted by $p^{(1)}(X_{1\delta}, X_{2\delta}, \cdots X_{M\delta})$, is given by the product of the discretised conditional density of $X_{k\delta}$ given $X_{(k-1)\delta}$ under $\mathbb{P}_1$, denoted by $p^{(1)}(X_{k\delta} \mid X_{(k-1)\delta})$, which is given by:

$$p^{(1)}(X_{k\delta} \mid X_{(k-1)\delta}) = \frac{1}{\sqrt{2\pi\sigma(X_{(k-1)\delta})^2\delta}} \exp\left(-\frac{\left(X_{k\delta} - X_{(k-1)\delta} - b^{(1)}(X_{(k-1)\delta})\delta\right)^2}{2\sigma(X_{(k-1)\delta})^2\delta}\right).$$

Therefore, we have:

$$\log\left(\frac{p^{(2)}(X_{k\delta} \mid X_{(k-1)\delta})}{p^{(1)}(X_{k\delta} \mid X_{(k-1)\delta})}\right) = \frac{(X_{k\delta} - X_{(k-1)\delta})(b^{(2)}(X_{(k-1)\delta}) - b^{(1)}(X_{(k-1)\delta}))\delta}{\sigma(X_{(k-1)\delta})^2\delta} + \frac{\left(b^{(1)}(X_{(k-1)\delta})^2 - b^{(2)}(X_{(k-1)\delta})^2\right)\delta^2}{2\sigma(X_{(k-1)\delta})^2\delta}.$$

Taking summation over k and $\delta \downarrow 0$ gives rise to the result. □

Example 8. Let $\mathbb{P}$ be the law induced by a $1-$dimensional Ito diffusion, and $\mathbb{W}$ be the one induced by a standard Brownian motion, that is $dX_t = \sigma(X_t)dW_t$. From Theorem 6, we have that:

$$\frac{d\mathbb{P}}{d\mathbb{W}}\left(X_{[0,t]}\right) = \exp\left(\int_0^t \frac{b(X_s)}{\sigma(X_s)}dX_s - \int_0^t \frac{b(X_s)^2}{2\sigma(X_s)^2}ds\right).$$

1.4.3 The Fokker–Planck equation

Let X_t be $d-$dimensional Ito diffusion. If the density $p_t(x)$ of the process exists, then p is given by the solution to the Fokker–Planck equation (1.9). Also if the stationary Fokker–Planck equation (1.11) holds, then the process is ergodic. Again, the Fokker–Planck equation of Ito diffusions is given by:

$$\begin{cases} \frac{\partial p}{\partial t} & = -\sum_{i=1}^d \frac{\partial}{\partial x_i}(b_i(x)p) + \frac{1}{2}\sum_{i,j=1}^d \frac{\partial^2}{\partial x_i \partial x_j}(\Sigma_{i,j}(x)p), \\ p(x,0) & = \mu_0(x), \end{cases} \tag{1.43}$$

where μ_0 is an initial density. Following Pavliotis (2014, Chapter 4), we define the following.

Definition 24. (Classical solution to the Fokker–Planck equation). A solution to the Fokker–Planck equation (1.43) is said to be *classical* if:

i) Transition density p of the process is twice continuously differentiable.

ii) For any $T > 0$, there exists a constant $c > 0$ such that $\|p_t(x)\|_{L^\infty(0,T)} \leq c\exp\left(a\|x\|^2\right)$, where L^∞ is the vector space of essentially bounded measurable functions with the essential supremum norm.

iii) $\lim_{t\to 0} p_t(x) = \mu_0(x)$.

Theorem 7. *(Pavliotis, 2014, Theorem 4.1) Assume that:*

i) There exists a constant $a > 0$ such that $\sum_{i,j=1}^d \Sigma_{i,j}(x)\xi_i\xi_j \geq a\|\xi\|^2$ for any $\xi \in \mathbb{R}^d$ uniformly in $x \in \mathbb{R}^d$.

ii) There exists a constant $M > 0$ such that $\|\Sigma(x)\| \leq M$, $\left\|\tilde{b}(x)\right\| \leq M(1 + \|x\|)$, $\|\tilde{c}(x)\| \leq M\left(1 + \|x\|^2\right)$ where $\tilde{b}(x) := -b_i(x) + \sum_{i=1}^d \frac{\partial \Sigma_{i,j}}{\partial x_j}$ and $\tilde{c}(x) := \frac{1}{2}\sum_{i,j=1}^d \frac{\partial^2}{\partial x_i \partial x_j}\Sigma_{i,j}(x) - \sum_{i=1}^d \frac{\partial b_i(x)}{\partial x_i}$.

Then there exists a unique classical solution to the Fokker–Planck equation so that the Ito diffusion is ergodic.

Next we consider the detailed balance condition for Ito diffusion, see 11 in the case of Markov chains. To do so, we first define:

$$J_i(p) := b_i(x)p - \frac{1}{2}\sum_{i=1}^{d}\frac{\partial}{\partial x_j}\left(\Sigma_{i,j}(x)p\right), \tag{1.44}$$

so that the Fokker–Planck equation (1.43) can be written as $\frac{\partial p}{\partial t} + \nabla J(p) = 0$, where $i-$th component of J is J_i in (1.44). Recall that the generator $\mathcal{L}$ of Ito diffusions is $\mathcal{L} = \sum_{i=1}^{d} b_i(x)\frac{\partial}{\partial x_i} + \frac{1}{2}\sum_{i,j=1}^{d}\Sigma_{i,j}(x)\frac{\partial^2}{\partial x_i \partial x_j}$. Therefore, we want to find a condition that makes $\mathcal{L}$ is a self-adjoint operator in $L_2(p)$. Assume that a Ito diffusion X_t is ergodic so that an invariant distribution p_t is the solution to the stationary Fokker-Plank equation (1.11):

$$\mathcal{L}^* p_t = 0,$$

where again $\mathcal{L}^*$ is the adjoint operator of $\mathcal{L}$, and notice that this implies $\nabla J(p_t) = 0$. Let $f, g \in L_2(p)$ be twice differentiable functions so that we want show that $\langle \mathcal{L}f, g\rangle_p = \langle f, \mathcal{L}^* g\rangle_p = \langle f, \mathcal{L}g\rangle$. Then it can be shown that:

$$\langle -\mathcal{L}f, g\rangle_p = \frac{1}{2}\langle \Sigma\nabla f, \nabla g\rangle_p + \left\langle f, p_t^{-1}\nabla g J(p_t)\right\rangle,$$

see Pavliotis (2014, Chapter 4) for details. From this $\mathcal{L}$ is a self-adjoint operator in $L_2(p)$ iff:

$$J(p_t) = 0, \tag{1.45}$$

and this is *the detailed balance condition* for Ito diffusions.

Proposition 10. *(Pavliotis, 2014, Proposition 4.5, Theorem 4.5) Let X_t be $d-$dimensional ergodic Ito diffusion with an invariant distribution p_t defined as $\mathcal{L}^* p_t = 0$. The the process is reversible iff (1.45) holds.*

1.4.4 Numerical approximation of SDEs

To simulate Ito diffusions, one needs to solve SDEs numerically since they do not have a closed-form solution in general. Given an Ito diffusion X_t on $[0, T]$ with $X_0 = x$, first one needs to divide the interval $[0, T]$ into M points such as $0 = t_0 < t_1 < \cdots t_M = T$ with a step size $\delta := \frac{T}{M} = t_i - t_{i-1}$. We write $Y_i = X_{t_i}$ with $t_i := \delta i$, that is $Y := \{Y_i; 0 \le i \le M\}$ is a discrete time approximation of $X := \{X_t; 0 \le t \le T\}$ on the grid $\{i\delta\}_{i=0}^{M}$. We then introduce the following two convergence criteria.

Definition 25. (Strong/Weak convergence). Let X_t be an Ito diffusion on $[0, T]$ with $X_0 = x$ and Y be its discrete time approximation on the grid $\{i\delta\}_{i=0}^{M}$. We say Y *converges strongly to X at time $T > 0$ with order $\gamma > 0$* if there exists a finite constant $C > 0$ and a positive constant δ_0 such that:

$$\mathbb{E}\left[\|X_T - Y_T\|\right] \le C\delta^{\gamma}, \tag{1.46}$$

for any $\delta \in (0, \delta_0)$. We say Y *converges weakly to* X *at time* $T > 0$ *with order* $\beta > 0$ if for any polynomial g, there exists a constant K_g and a positive constant δ_0 such that

$$|\mathbb{E}[g(X_T)] - \mathbb{E}[g(Y_T)]| \leq K_g \delta^\beta, \tag{1.47}$$

for any $\delta \in (0, \delta_0)$ provided that these functionals exist.

Complete reviews of the numerical solutions of SDEs can be found in Platen (1999); Higham (2001). In particular, we focus on the *Euler-Maruyama scheme*. The Euler-Maruyama scheme of X is given by the recursive equation:

$$Y_{i+1} = Y_i + b(Y_i)\delta + \sigma(Y_i)\delta W_{i+1}, \tag{1.48}$$

for $0 \leq i \leq M-1$ where $Y_0 = X_0$, $\delta := \frac{T}{M}$, $\delta W_i := W_{t_i} - W_{t_{i-1}}$ and thus $\delta W_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \delta)$. Then the corresponding continuous time approximation of X_t is as follows:

$$\hat{Y}_t := Y_i + \int_{t_i}^{t} b(X_{t_i})ds + \int_{t_i}^{t} \sigma(X_{t_i})dW_s, \tag{1.49}$$

so that $\hat{Y}_{t_i} = Y_i$. The following theorem provides that the strong convergence order of the Euler-Maruyama scheme is $\frac{1}{2}$, we omit the proof.

Theorem 8. *Mao (2007, Theorem 2.7.3) Let X_t be an Ito diffusion on $[0,T]$ with $X_0 = x$ and $\hat{Y}_t$ be its continuous time approximation (1.49). Then for any $p \geq 1$, under 1.33, we have that:*

$$\mathbb{E}\left[\sup_{0\leq t\leq T} \left\|X_t - \hat{Y}_t\right\|^p\right] \leq \frac{C(p,T)}{N^{p/2}},$$

where $C(p,T)$ is a constant depending on p, T.

We also note that (1.48) induces the Markov chain $x \mapsto x + \delta b(x) + \sqrt{\delta}\sigma(x)W$, $W \sim \mathcal{N}(0, I_d)$. Assume that:

$$\lim_{|x|\to\infty} \frac{2xb(x) + \text{trace}\left(\sigma^\top(x)\sigma(x)\right) + \delta \mid b(x) \mid^2}{\mid x \mid^2} \leq 0. \tag{1.50}$$

If one takes $V(x) = \mid x \mid^2$, then (1.50) impels that there exist $b, \gamma \in (0, \infty)$ such that $PV = 2\delta xb(x) + trace\left(\sigma^\top(x)\sigma(x)\right)\delta + \delta^2 \mid b(x) \mid^2 \leq b - \gamma V$ so that we have $PV \leq b + \lambda V$ with $\lambda = 1 - \gamma < 1$ thus (18) holds. Also for any $r \in (0, \infty)$, we have that:

$$\begin{aligned}\Delta^L(P, \{x \in \mathbb{R}^d; \mid x \mid^2 \leq r\}) &= \sup_{|x|\leq\sqrt{r}} \sup_{|y|\leq\sqrt{r}} \left\|\mathcal{N}\left(x + \delta b(x), \delta\sigma(x)\sigma(x)^\top\right) - \mathcal{N}\left(y + \delta b(y), \delta\sigma(y)\sigma(y)^\top\right)\right\|_{TV}, \\ &\leq 1,\end{aligned}$$

thus we obtain the following, see Eberle (2020) for instance.

Proposition 11. *Let X_t be an Ito diffusion on $[0,T]$ with $X_0 = x$ and Y be its Euler-Maruyama*

approximation in(1.48). *Assume that* (1.50) *holds. Then the Markov chain induced by the Euler-Maruyama approximation of* X_t *is geometrically ergodic.*

2 Markov chain Monte Carlo methods

2.1 Introduction

Suppose that one wants to evaluate an integral of the form of $\pi(f) := \int f(x)\pi(dx)$ for $f \in \mathcal{B}_b(E)$ and $\pi \in \mathcal{P}(E)$. In Bayesian statistics, $\pi(dx)$ is often (maybe always) a posterior distribution and thus such integral will be the posterior expectation of f. If one could generate random numbers directly from the target $\pi(dx)$, a natural approximation of $\pi(f)$ would be:

$$\hat{\pi}(f) := \frac{1}{N}\sum_{i=1}^{N} f(x^{(i)}),$$

where $x^{(i)} \overset{i.i.d.}{\sim} \pi(dx)$. This class of approximation methods is called a *Monte Carlo method*, we refer to Robert and Casella (2013) for a general reference of Monte Carlo methods. Suppose that one has $\tilde{\pi} \in \mathcal{S}(E)$ where, again, $\mathcal{S}(E)$ denotes the set of all finite signed measures over the measurable space $(E, \mathcal{E})$. In practice, $\tilde{\pi}(dx)$ will correspond to a posterior density w.r.t. the Lebesgue measure known up to a normalising constant. That is, $\tilde{\pi}(dx)$ will be the product of a likelihood function and a (non-conjugate) prior distribution. In this case, one cannot use the vanilla Monte Carlo any more.

Instead of sampling directly from a target distribution, *Markov chain Monte Carlo* (MCMC) methods generate (dependent) samples from a Markov chain which is constructed by users. That is, the main objective of MCMC is that, given a target distribution, one has to construct an ergodic Markov chain which has the target distribution as an invariant distribution.

The main objective of this section is to provide some basic and detailed results of a variety of MCMC methods which will appear implicitly and explicitly in the rest of the thesis. We refer to Roberts and Rosenthal (2004) as a general reference of MCMC methods.

2.2 Metropolis–Hastings

Following closely Tierney (1994, 1998), we begin with the *Metropolis-Hastings algorithm*. Most MCMC methods are (partially or fully) based on the Metropolis-Hastings algorithm, or can be considered as a special case of it. Let $\pi(dx)$ be a target distribution on $(E, \mathcal{E})$ which is potentially known up to a normalizing constant. In order to define the Metropolis–Hastings kernel for $\pi(dx)$, one has to specify a proposal Markov kernel $Q(x, dy)$ on $(E, \mathcal{E})$, which can be admitting a density q w.r.t. the Lebesgue measure v, that is $Q(x, dy) = q(x, y)v(dy)$. Based on $Q(x, dy)$, we want to construct the Markov kernel $P(x, dy)$ which satisfies:

$$\int_A \pi(dx)P(x, B) = \int_B \pi(dx)P(x, A), \tag{2.1}$$

for any $A, B \in \mathcal{E}$, or equivalently $\pi(dx)P(x,dy) = \pi(dy)P(y,dx)$ on the product space $(E \times E, \mathcal{E} \otimes \mathcal{E})$. This is because taking $A = E$ and (2.1) give rise to:

$$\begin{aligned}\int_E \pi(dx)P(x,B) &= \int_B \pi(dx)P(x,E), \\ &= \pi(B),\end{aligned}$$

which implies that $\pi P = \pi$. Therefore, the condition (2.1) is a sufficient condition that a Markov kernel $P(x,dy)$ leaves $\pi(dx)$ invariant. Recall that the condition (2.1) is often called the *detailed balance condition*. Also, notice that (2.1) implies:

$$\mathbb{P}\left(X_n \in A, X_{n+1} \in B\right) = \mathbb{P}\left(X_n \in B, X_{n+1} \in A\right),$$

thus such Markov chain is reversible. Then consider the *Metropolis–Hastings kernel*:

$$P(x,dy) = Q(x,dy)\alpha(x,y) + \delta_x(dy)\left(1 - \bar{\alpha}(x)\right), \tag{2.2}$$

$$\bar{\alpha}(x) := \int \alpha(x,y)Q(x,dy). \tag{2.3}$$

where $\alpha(x,y)$ is a measurable function such that $\alpha(x,y) : E \times E \to [0,1]$. Thus, (2.2) can be algorithmically understood as:

i) Given x, propose a move y via $Q(x,dy)$.

ii) Accept y w.p. $\alpha(x,y)$.

iii) Otherwise, stay at x.

Then, it turns out that (2.2) satisfies (2.1) iff

$$\pi(dx)Q(x,dy)\alpha(x,y) = \pi(dy)Q(y,dx)\alpha(y,x) \tag{2.4}$$

holds. To obtain such well-defined $\alpha(x,y)$, let $\mu(dx,dy) := \pi(dx)Q(x,dy)$ and $\mu^\top(dx,dy) := \mu(dy,dx) = \pi(dy)Q(y,dx)$. Then it can be shown that there exists a set $C \in \mathcal{E} \otimes \mathcal{E}$ such that μ and $\mu^\top$ are mutually absolutely continuous on C and mutually singular on C^c. Then we write μ_C and $\mu_C^\top$ as the restrictions of μ and $\mu^\top$ on C. From the Radon–Nikodým theorem, we have that (Tierney, 1998, Proposition1):

$$r(x,y) := \frac{\mu_C(dx,dy)}{\mu_C^\top(dx,dy)}, \tag{2.5}$$

such that $0 < r(x,y) < \infty$ and $r(x,y) = \frac{1}{r(y,x)}$ for all $x, y \in E$. In practice, the set C can be understood as the state such that moves from x to y and from y to x are both possible in the Markov chain with some initial distribution and the transition kernel $Q(x,dy)$. Then the detailed condition in (2.4) can be alternatively expressed as:

$$\mu(dx,dy)\alpha(x,y) = \mu^\top(dx,dy)\alpha(y,x). \tag{2.6}$$

From this, it can be shown that (2.2) satisfies (2.1) iff

$$\alpha(y,x) = \frac{\mu(dx,dy)}{\mu^{\top}(dx,dy)}\alpha(x,y) = r(x,y)\alpha(x,y) \tag{2.7}$$

holds w.p.1 on C, see e.g. Tierney (1998, Theorem2). Then for $x, y \in E$, consider the following *acceptance probability function* α_{MH} :

$$\alpha_{MH}(x,y) := \begin{cases} \min\{1, r(y,x)\}, \ if \ (x,y) \in C, \\ 0, \ otherwise. \end{cases} \tag{2.8}$$

Then for $x, y \in C$, we have that:

$$\begin{aligned} r(x,y)\alpha_{MH}(x,y) &= \min\{r(x,y), r(x,y)r(y,x)\}, \\ &= \min\{r(x,y), 1\} = \alpha_{MH}(y,x), \end{aligned}$$

thus (2.2) satisfies (2.1) with $\alpha_{MH}(x,y)$. If we assume that there exists a common dominating measure ν such that $\pi(dx) = \pi(x)\nu(dx)$ and $Q(x,dy) = q(y \mid x)\nu(dy)$ hold. Then we can take:

$$C = \{(x,y) : \pi(x)q(y \mid x) > 0, \pi(y)q(x \mid y) > 0\},$$

and:

$$r(x,y) = \frac{\pi(x)q(y \mid x)}{\pi(y)q(x \mid y)}. \tag{2.9}$$

Then the following theorem immediately follows from the construction.

Theorem 9. *Consider the Metropolis–Hastings kernel* (2.2) *with acceptance probability function* α_{MH} *defined in* (2.8)*. That is:*

$$\begin{aligned} P(x,dy) &= Q(x,dy)\alpha_{MH}(x,y) + \delta_x(dy)\left(1 - \bar{\alpha}_{MH}(x)\right), \\ \bar{\alpha}_{MH}(x) &:= \int \alpha_{MH}(x,y)Q(x,dy). \end{aligned}$$

Then we have that $\pi(dx)P(x,dy) = \pi(dy)P(y,dx)$.

We can informally summarise above discussions as follows under the assumption that there exists a common dominating measure ν such that $\pi(dx) = \pi(x)\nu(dx)$ and $Q(x,dy) = q(y \mid x)\nu(dy)$ hold.

Algorithm 1 Metropolis–Hastings

i) Given x, propose a move y via $q(\cdot \mid x)$.

ii) Set $x = y$ w.p. $\min\left\{1, \frac{\pi(y)q(x|y)}{\pi(x)q(y|x)}\right\}$.

iii) Repeat step 1 and 2 enough times.

Remark 3. *Notice that* $\frac{\pi(y)q(x|y)}{\pi(x)q(y|x)}$ *does not depend on the normalising constant of the density* $\pi(x)$. *Indeed,* $\frac{\pi(y)}{\pi(x)} = \frac{\tilde{\pi}(y)/Z}{\tilde{\pi}(x)/Z} = \frac{\tilde{\pi}(y)}{\tilde{\pi}(x)}$ *where* $\tilde{\pi} \in \mathcal{S}(E)$ *and* Z *is the corresponding normalising constant.*

Example 9. *Independent type Metropolis-Hastings.*

Set a proposal density such as $q(y \mid x) = q(y)$. Then (2.9) becomes $r(x, y) = \frac{\pi(y)q(x)}{\pi(x)q(y)}$. The Metropolis–Hastings with this choice of the proposal density is often called the *independent type Metropolis-Hastings*.

Example 10. *Random Walk Metropolis.*

Set a proposal density such as $q(y \mid x) = q(x \mid y)$. The most common choice might be $q(y \mid x) = q(x - y) = q(y - x)$. Then (2.9) becomes $r(x, y) = \frac{\pi(y)}{\pi(x)}$. The Metropolis–Hastings with this choice of the proposal density is often called the *random walk Metropolis-Hastings*. A typical choice of such q is $y = x + \epsilon$ where $\epsilon \overset{i.i.d.}{\sim} \mathcal{N}\left(0, \sigma^2 I\right)$ with $\sigma^2 > 0$.

Although there are many valid acceptance probability functions satisfying the condition (2.6), critically, the Metropolis–Hastings kernel (2.2) with (2.8) is *optimal in terms of the asymptotic variance*. To see this, again define $L^2(\pi)$ such that:

$$L^2(\pi) := \left\{ x \in E : \int_E f(x)\pi(dx) < \infty \right\},$$

equipped with the inner product $\langle f, g \rangle := \int_E f(x)g(x)\pi(dx)$. Without loss of generality, we can restrict our attention to the space $L_0^2(\pi) \subset L^2(\pi)$, where:

$$L_0^2(\pi) := \left\{ f \in L^2(\pi) : \int_E f(x)\pi(dx) = 0 \right\}.$$

Notice that $\langle f, g \rangle = \mathrm{Cov}(f, g)$ for $f, g \in L_0^2(E)$. Suppose that a Markov kernel $P(x, dy)$ satisfies (2.1) w.r.t. $\pi(dx)$. In this case, we have that:

$$\begin{aligned}\langle Pf, g \rangle &= \int\int P(x, dy)f(y)g(x)\pi(dx) = \int\int P(y, dx)f(y)g(x)\pi(dy), \\ &= \int\int P(x, dy)f(x)g(y)\pi(dx) = \langle f, Pg \rangle,\end{aligned}$$

implying P is self-adjoint on $L^2(\pi)$, again. For $f \in L^2(\pi)$, the *Dirichlet form* for the Markov kernel P is given by:

$$\begin{aligned}\mathcal{E}_P(f) &:= \langle f, f \rangle - \langle f, Pf \rangle, \\ &= \int f(x)^2\pi(dx) - \int\int f(x)P(x, dy)f(y)\pi(dx), \\ &= \int\int f(x)\left[f(x) - f(y)\right]\pi(dx)P(x, dy), \\ &= \int\int f(y)\left[f(y) - f(x)\right]\pi(dx)P(x, dy), \\ &= \frac{1}{2}\int\int \left[f(y) - f(x)\right]^2\pi(dx)P(x, dy). \end{aligned} \tag{2.10}$$

Next, define $\hat{f}_N := \frac{1}{N}\sum_{i=1}^{N} P^{i-1}f(x)$ for $f \in L^2(\pi)$. Then the asymptotic variance $Var_\pi(f,P)$ is defined as:

$$Var_\pi(f,P) := \lim_{N\to\infty} Var(\hat{f}_N) = \lim_{N\to\infty} Var\left(\sqrt{N}\hat{f}_N\right). \tag{2.11}$$

Suppose that $f \in L_0^2(\pi)$ and $X_0 \sim \pi$. Then it can be shown that:

$$Var_\pi(f,P) = \left\langle f, (I+P)(I-P)^{-1}f \right\rangle, \tag{2.12}$$

and:

$$\langle f, (I-P)f\rangle = \sup_{g\in L_0^2(E,\pi)} 2\langle f,g\rangle - \mathcal{E}_P(g),$$

see Sherlock (2018) for details. As a result, for $f \in L_0^2(\pi)$ and $X_0 \sim \pi$, we can show that:

$$Var_\pi(f,P) = \sup_{g\in L_0^2(E,\pi)} 4\langle f,g\rangle - \mathcal{E}_P(g) - \langle f,f\rangle, \tag{2.13}$$

here note that $\langle f,f\rangle = \left\langle f,(I-P)(I-P)^{-1}f\right\rangle$ so that $\left\langle f,(I+P)(I-P)^{-1}f\right\rangle = 2\left\langle f,(I-P)^{-1}f\right\rangle - \langle f,f\rangle$. Suppose that one has two $\pi-$invariant Markov kernels P_1 and P_2. Then we immediately have the following.

Theorem 10. *Tierney (1998). Suppose that one has two $\pi-$reversible Markov kernels P_1 and P_2 such that for any $g \in L_0^2(\pi)$, $\mathcal{E}_{P_1}(g) \geq \mathcal{E}_{P_2}(g)$. Then $Var_\pi(f,P_1) \leq Var_\pi(f,P_2)$ holds.*

Now consider the Metropolis–Hastings kernel. From (2.10), the Dirichlet form of the Markov kernels is:

$$\begin{aligned}\mathcal{E}_P(f) &= \frac{1}{2}\int\int [f(y)-f(x)]^2 \pi(dx)\left(Q(x,dy)\alpha(x,y) + \delta_x(dy)\left(1-\bar{\alpha}(x)\right)\right),\\ &= \frac{1}{2}\int\int [f(y)-f(x)]^2 \pi(dx)Q(x,dy)\alpha(x,y).\end{aligned}$$

Also, it is clear to see that:

$$\alpha(x,y) = r(y,x)\alpha(y,x) \leq \min\{1, r(y,x)\} = \alpha_{MH}(x,y)$$

holds for any $\alpha(x,y)$ satisfying (2.7). Then write P_{MH} for the Metropolis–Hastings kernel with (2.8) and P_α for any $\pi-$invariant reversible Markov kernel with any $\alpha(x,y)$ satisfying (2.7). For fixed $Q(x,dy)$, we clearly have that:

$$\mathcal{E}_{P_{MH}}(g) \geq \mathcal{E}_{P_\alpha}(g), \tag{2.14}$$

for any $g \in L_0^2(\pi)$. This leads to the following.

Theorem 11. *Tierney (1998). Let $Q(x,dy)$ be fixed. Then amongst reversible Markov kernels P of*

the form:

$$P(x, dy) = Q(x, dy)\alpha(x, y) + \delta_x(dy)\left(1 - \bar{\alpha}(x)\right),$$
$$\bar{\alpha}(x) := \int \alpha(x, y)Q(x, dy),$$

where $\alpha(x, y)$ *satisfying* (2.7)*, the one minimizing* $Var_\pi(f, P)$ *for any* $f \in L^2(E, \pi)$ *is* $P(x, dy) = Q(x, dy)\alpha_{MH}(x, y) + \delta_x(dy)\left(1 - \bar{\alpha}_{MH}(x)\right)$.

Proof. This follows from (2.14) and Theorem 10. □

Remark 4. *Theorem 11 tells us only about the asymptotic variance of* $\pi-$*reversible Markov kernels. Thus, it does not tell us anything about non-reversible Markov kernels, or about non-asymptotic variance.*

Next, we consider the ergodicity of the Metropolis-Hastings algorithm. Consider first the independent type Metropolis-Hastings (Example 9). Assume that for a target density π and a proposal density q, there exists $\epsilon > 0$ such that:

$$\inf_{x \in E} \frac{q(x)}{\pi(x)} \geq \epsilon. \tag{2.15}$$

Then we have the following.

Proposition 12. *Assume that the constant defined in* (2.15) *exists. Then the independent type Metropolis-Hastings is uniformly ergodic.*

Proof. Let $P(x, \cdot)$ be the Markov kernel induced by the independent type Metropolis-Hastings. Then for any $x \in E$ and $A \in \mathcal{E}$, we have that:

$$\begin{aligned} P(x, A) &\geq \int_A q(y) \min\left\{1, \frac{\pi(y)q(x)}{\pi(x)q(y)}\right\} dy, \\ &= \int_A \pi(y) \min\left\{\frac{q(y)}{\pi(y)}, \frac{q(x)}{\pi(x)}\right\} dy, \\ &\geq \epsilon \int_A \pi(y)dy = \epsilon\pi(A). \end{aligned}$$

Notice that this ϵ is necessarily $\epsilon \in (0, 1]$ since $1 = P(x, E) \geq \epsilon\pi(E) = \epsilon$. Thus the Markov kernel induced by the algorithm satisfies the Doeblin condition, and the result follows. □

We also consider the random walk Metropolis-Hastings (Example 10). Critically, Roberts and Tweedie (1996b, Proposition 3.1) show that, in the case of the Hastings-Metropolis algorithm, a Markov kernel is geometrically ergodic *iff* there exists a real-valued function $V > 1$ such that:

$$\limsup_{|x| \to \infty} \frac{PV(y)}{V(x)} < 1, \tag{2.16}$$

where $PV(y) := \int P(x, dy)V(y)$ again. Recall that the Markov kernel induced by the Hastings-Metropolis algorithm is given by $P(x, dy) = Q(x, dy)\alpha_{MH}(x, y) + \delta_x(dy)\left(1 - \int \alpha_{MH}(x, y)Q(x, dy)\right)$ so

that we have:

$$\begin{aligned}\frac{PV(y)}{V(x)} &= \int \frac{V(y)}{V(x)} Q(x,dy)\alpha_{MH}(x,y) + \int \frac{V(y)}{V(x)} a(x)\delta_x(dy) = \int \frac{V(y)}{V(x)} Q(x,dy)\alpha_{MH}(x,y) + a(x) \\ &= \int \left[\frac{V(y)}{V(x)} - 1\right] Q(x,dy)\alpha_{MH}(x,y) + 1,\end{aligned}$$

where we have defined $a(x) := 1 - \int \alpha_{MH}(x,y)Q(x,dy)$. Thus, (2.16) becomes:

$$\limsup_{|x|\to\infty} \int \left[\frac{V(y)}{V(x)} - 1\right] Q(x,dy)\alpha_{MH}(x,y) < 0. \tag{2.17}$$

To utilise the criterion in (2.17), we need to restrict the class of a target density. Let $p(\cdot)$ be a continuous and positive density over $\mathbb{R}$. Then $p(\cdot)$ is said to be *log-concave in the tails* if there exists $a > 0$ and some x' such that for all $y \geq x \geq x'$:

$$\log p(x) - \log p(y) \geq a(y - x),$$

and also for any $y \leq x \leq -x'$:

$$\log p(x) - \log p(y) \geq a(x - y),$$

hold.

Proposition 13. *Let P be the Markov kernel induced by the random walk Metropolis-Hastings. Assume that $E = \mathbb{R}$ and a target density $\pi(\cdot)$ is log-concave in tails. Then P is geometrically ergodic.*

Proof. We adopt the proof of Roberts and Tweedie (1996b); Mengersen and Tweedie (1996). For the sake of simplicity, assume that $q(y \mid x) = q(x - y) = q(y - x)$. Take $V(x) = \exp(s \mid x \mid)$ for $0 < s < a$. Also notice that we can the integral in (2.17) as:

$$\begin{aligned} I &:= \int \left[\frac{V(y)}{V(x)} - 1\right] Q(x,dy)\alpha_{MH}(x,y) \\ &= \underbrace{\int_{-\infty}^{0} \left[e^{s(|y|-x)} - 1\right] \alpha_{MH}(x,y)Q(x,dy)}_{A} + \underbrace{\int_{0}^{x} \left[e^{s(y-x)} - 1\right] \alpha_{MH}(x,y)Q(x,dy)}_{B} \\ &+ \underbrace{\int_{x}^{2x} \left[e^{s(y-x)} - 1\right] \alpha_{MH}(x,y)Q(x,dy)}_{C} + \underbrace{\int_{2x}^{\infty} \left[e^{s(y-x)} - 1\right] \alpha_{MH}(x,y)Q(x,dy)}_{D}. \end{aligned}$$

Then for $0 < s < a$ and $x \geq x'$, the assumptions imply $\frac{\pi(y)}{\pi(x)} \leq \exp(-a(\mid y - x \mid))$ so that:

$$\begin{aligned} D &= \int_{2x}^{\infty} \left[e^{s(y-x)} - 1\right] \frac{\pi(y)}{\pi(x)} Q(x,dy) \\ &\leq \exp(2x(s-a))\, Q(x,(2x,\infty)) \to 0, \end{aligned}$$

by taking $x' \to \infty$. Also we have that:

$$\begin{aligned} A &\leq \int_{-\infty}^{-x} \left[e^{s(|y|-x)} - 1\right] e^{-a(|y|-x)} Q(x, dy) + \int_{-x}^{0} \left[e^{s(|y|-x)} - 1\right] Q(x, dy), \\ &\leq Q(x, (-x, -\infty)) + \int_{-x}^{0} \left[e^{s(|y|-x)} - 1\right] Q(x, dy). \end{aligned}$$

Notice that the second integral on the right hand side is strictly negative for any x so that $A \to 0$ as $x' \to \infty$. As for B and C, recall that we have assumed $q(x-y) = q(y-x)$. Therefore, we can bound these two terms by:

$$\int_0^x \left[e^{-sz} - 1 + e^{(s-a)z} + e^{-az}\right] q(z) dz = -\int_0^x (1 - e^{(s-a)z})(1 - e^{-sz}) q(z) dz,$$

whose integrand is positive and increasing in z so that strictly negative, thus $I \to < 0$ as $x' \to \infty$. The symmetricity of the log-concave assumption ensures that the same argument holds if $x' \to -\infty$ so that the claim follows from Roberts and Tweedie (1996b, Proposition 3.1). □

To get Proposition 13, we need to assume that target distributions are log-concave in the tails. Indeed, Mengersen and Tweedie (1996, Theorem 3.3) show the following necessary condition for geometric ergodicity.

Theorem 12. *Mengersen and Tweedie (1996). Let P be the Markov kernel induced by the random walk Metropolis-Hastings. Let $\pi \in \mathcal{P}(E)$ be a target. If P is geometrically ergodic then there exists $s > 0$ such that:*

$$\int e^{s|x|} \pi(dx) < \infty.$$

Theorem 12 essentially implies that the random walk Metropolis-Hastings cannot be geometrically ergodic for targets with heavy tails.

Again, consider the random walk Metropolis-Hastings with the proposal density $q(y \mid x) = \mathcal{N}\left(y; x, \sigma^2 I\right)$. Given this choice, another natural question would be the optimal choice of the parameter σ^2. In this context, Roberts et al. (1997) study the targets of the form $\pi(x) = \prod_{i=1}^{d} f(x_i)$ with Gaussian jump proposals $y^{(d)} \sim \mathcal{N}\left(0, \sigma^2(d) I_d\right)$ here we have used d to emphasise the dependency on dimension. As d increases, the number of proposed moves obviously increases so that the random walk Metropolis-Hastings becomes more likely to propose unreasonable moves. As a consequence, the acceptance probability $\alpha_{MH}(x, y)$ (2.8) might degenerate into 0 as $d \uparrow \infty$. To overcome this problem, Roberts et al. (1997) consider *scaling* the parameter $\sigma^2(d)$:

$$\sigma^2(d) := \frac{\ell^2}{d}, \tag{2.18}$$

where ℓ is a positive constant. Clearly, (2.18) is a decreasing function of dimension d. The key idea of

Roberts et al. (1997) is that they consider the following Markov process:

$$Z_1^{(d)}(t) := X_{1\lfloor td \rfloor}^{(d)},$$

where $\lfloor \cdot \rfloor$ is the floor function and thus $X_{1\lfloor td \rfloor}^{(d)}$ is the first component of the chain after iteration $\lfloor td \rfloor$. Therefore, the algorithm proposes a move every $1/d$ time step. This process allows us to consider asymptotic behaviour of the chain with $d \uparrow \infty$ since that rescales the time between each step as well. Under some analytical conditions on $\pi(x)$, one can show that :

$$Z_1^{(d)}(t) \to Z(t),$$

where the convergence is in distribution, and $Z(t)$ satisfies the Langevin equation such that:

$$\begin{aligned} dZ_t &= \frac{1}{2} h(\ell) \nabla \log f(Z_t) + \sqrt{h(\ell)} dW_t, \\ h(\ell) &:= 2\ell^2 \Phi\left(-\frac{1}{2}\ell I^{1/2}\right), \\ I &:= \mathbb{E}\left[(\nabla \log f)^2\right]. \end{aligned} \tag{2.19}$$

Here, $h(\ell)$ is sometimes called a *speed* measure for the diffusion process in the sense that $Z(t)$ can be expressed as a sped up version of the process $U(t)$, i.e., $Z(t) = U(h(\ell)t)$ where $=$ is in distribution sense, where $U(t)$ is given by:

$$dU_t := \frac{1}{2} \nabla \log f(U_t) dt + dW_t.$$

Indeed, it can be shown that setting $ds = h(\ell)dt$ gives rise to $dU_t = dZ_t$, again $=$ is in distribution. Consider the *expected acceptance rate* in $d-$dimensions:

$$\begin{aligned} \alpha_{MH}(d, \ell) &:= \int \int \pi(x^{(d)}) \alpha_{MH}(x^{(d)}, y^{(d)}) q(y^{(d)} \mid x^{(d)}) dx^{(d)} dy^{(d)} \\ &= \int \int \min\left\{\pi(x^{(d)}) q(y^{(d)} \mid x^{(d)}), \pi(y^{(d)}) q(x^{(d)} \mid y^{(d)})\right\} dx^{(d)} dy^{(d)}. \end{aligned} \tag{2.20}$$

Then the following theorem says that $\alpha_{MH}(d, \ell)$ converges to $2\Phi\left(-\frac{1}{2}\ell I^{1/2}\right)$ so maximising $h(\ell)$ gives rise to the optimal choice of $\sigma^2(d)$ in terms of the acceptance probability.

Theorem 13. *Roberts et al. (1997, Theorem 1.1, Corollary 1 .2). Under some regular conditions on* $\pi(x)$, $X_{1\lfloor td \rfloor}^{(d)}$ *converges to* $Z(t)$ *defined in* (2.19) *in distribution as* $d \uparrow \infty$. *Also the* expected acceptance rate *in* $d-$*dimensions* $\alpha_{MH}(d, \ell)$ *converges to* $\alpha(\ell) := 2\Phi\left(-\frac{1}{2}\ell I^{1/2}\right)$ *as* $d \uparrow \infty$. $h(\ell)$ *is maximised at the unique value* $\hat{\ell} = 2.38/I^{1/2}$ *for which* $\alpha(\hat{\ell}) = 0.234$ *and* $h(\hat{\ell}) = 1.3/I$.

2.3 Reversible jump MCMC

The *reversible jump MCMC* (RJMCMC) algorithm is a trans-dimensional version of the Metropolis-Hastings algorithm, developed in Green (1995). Due to its nature, RJMCMC has been used especially in the context of Bayesian model selection and mixture models, for instance, see Hastie and Green

(2012); Robert and Casella (2013, Chapter 11) for details. For the sake of simplicity, we bravely confine our attention to Bayesian model selection.

We first introduce Bayesian model selection. Assume that one has a countable set of k parametric models, denoted by $\mathcal{M} := \{\mathcal{M}_k\}_{k\in\mathcal{K}}$ associated with a collection of likelihood functions $p(y \mid \theta_k, \mathcal{M}_k)$ where y is data, $\theta_k \in \Theta_k$ is the parameter and the parameter space respectively. Also, one has to specify a collection of priors on the parameters θ_k denoted by $\pi(\theta_k)$ and ones on models denoted by $\pi(\mathcal{M}_k)$. In general, $\pi(\theta_k)$ is a density w.r.t. the Lebesgue measure on Θ_k and $\pi(\mathcal{M}_k)$ is a one w.r.t. the counting measure on $\mathcal{M}_k$. Set $\Theta := \cup_k \Theta_k \times \{\mathcal{M}_k\}$. The *posterior model probability* of $\mathcal{M}_k$ given y can be obtained as:

$$\Pi(\mathcal{M}_k \mid y) = \frac{\pi(\mathcal{M}_k)\int p(y \mid \theta_k, \mathcal{M}_k)\pi(\theta_k)}{\sum_{l\in\mathcal{K}} \pi(\mathcal{M}_l)\int p(y \mid \theta_l, \mathcal{M}_l)\pi(\theta_l)}, \tag{2.21}$$

where we have assumed that θ_k and $\mathcal{M}_k$ are independent for each $k \in \mathcal{K}$.

The main difficulty to explore on the space Θ, or equivalently on the joint posterior $\Pi(\mathcal{M}_k, \theta_k) \propto p(y \mid \theta_k, \mathcal{M}_k)\pi(\theta_k)\pi(\mathcal{M}_k)$ is that $\dim(\Theta_k)$ and $\dim(\Theta_{k'})$ may differ, where $\dim(x)$ denotes the dimension of x. Consider a move from $x_k := (\mathcal{M}_k, \theta_k)$ to $x_{k'} := (\mathcal{M}_{k'}, \theta_{k'})$. Then, what we want to construct is that the Markov kernel $P(x, dx')$ satisfies the following detailed balance condition (2.4):

$$\int_{(x,x')\in A\times B} \Pi(dx)P(x, dx') = \int_{(x,x')\in A\times B} \Pi(dx')P(x', dx), \tag{2.22}$$

for any $A, B \in \Theta$. To construct such a kernel, we again use the Metropolis–Hastings kernel (2.2), that is (2.22) becomes:

$$\int_{(x,x')\in A\times B} \Pi(dx)Q(x, dx')\alpha(x, x') = \int_{(x,x')\in A\times B} \Pi(dx')Q(x', dx)\alpha(x', x), \tag{2.23}$$

Again, as we discussed (see also Tierney, 1998; Green, 1995), it can be shown that $\Pi(dx)Q(x, dx')$ is dominated by a symmetric measure μ on $\Theta \times \Theta$ with the corresponding density f. Intuitively, this means that jumps are limited to moves from $\mathcal{M}_k$ to close models in the sense that their dimensions are close (but slightly different), might be nested. As a result, (2.6) becomes:

$$\int_{(x,x')\in A\times B} f(x, x')\alpha(x, x')\mu(dx, dx') = \int_{(x,x')\in A\times B} f(x', x)\alpha(x', x)\mu(dx', dx). \tag{2.24}$$

Recall (see (2.7)) that (2.24) holds iff $\alpha(x, x') = \frac{f(x',x)}{f(x,x')}$, and this leads to the following *Green's ratio*:

$$\alpha_G(x, x') := \frac{\Pi(dx')Q(x', dx)}{\Pi(dx)Q(x, dx')}, \tag{2.25}$$

clearly the Metropolis–Hastings kernel with $\alpha_G(x, x')$ leaves $\Pi(dx)$ invariant due to Theorem 9.

Notice that although (2.25) is mathematically well defined, a constructive representation of $\mu(dx, dx')$ is still not clear. Then, the idea of Green (1995) is to impose a *dimension matching condition* in the sense that there exists transformation from θ_k to $\theta_{k'}$ such that it is a diffeomorphism (the transformation and its inverse mapping are differentiable). Let d_k and $d_{k'}$ be the dimensions of θ_k and $\theta_{k'}$. Then

one first simulates random variables $u_{k\to k'} \in \mathcal{U}_{k\to k'} \subset \mathbb{R}^{\ell_{k\to k'}}$ and set $(\theta_{k'}, u_{k'\to k}) = t_{k\to k'}(\theta_k, u_{k\to k'})$ where $t : \Theta_k \times \mathcal{U}_{k\to k'} \to \Theta_{k'} \times \mathcal{U}_{k'\to k}$ is a diffeomorphism, that is its inverse mapping is $t_{k\to k'} := t^{-1}_{k\to k'} : \Theta_{k'} \times \mathcal{U}_{k'\to k} \to \Theta_k \times \mathcal{U}_{k\to k'}$, where $u_{k'\to k} \in \mathcal{U}_{k'\to k} \subset \mathbb{R}^{\ell_{k'\to k}}$, and thus $d_k + \ell_{k\to k'} = d_{k'} + \ell_{k'\to k}$. Therefore, RCMCMC extends the spaces Θ_k and $\Theta_{k'}$ to the augmented spaces $\Theta_k \times \mathcal{U}_{k\to k'}$ and $\Theta_{k'} \times \mathcal{U}_{k'\to k}$. Let $g_{k\to k'}(u_{k\to k'})$ and $g_{k'\to k}(u_{k'\to k})$ be the densities of $u_{k\to k'}$ and $u_{k'\to k}$. As a result, now our augmented target densities are $\Pi(\mathcal{M}_k, \theta_k)g_{k\to k'}(u_{k\to k'})$ and $\Pi(\mathcal{M}_{k'}, \theta_{k'})g_{k'\to k}(u_{k'\to k})$. Define the Jacobian of the transformation $t_{k\to k'}$:

$$J_{k\to k'}(\theta_k, u_{k\to k'}) := \det \left| \left(\frac{\partial t_{k\to k'}(\theta_k, u_{k\to k'})}{\partial(\theta_k, u_{k\to k'})} \right) \right|, \tag{2.26}$$

and now we are ready to summarise the discussion as the following algorithm.

Algorithm 2 Reversible Jump MCMC (Green, 1995).

i) Given $(\mathcal{M}_k, \theta_k)$, sample $\mathcal{M}_{k'} \mid \mathcal{M}_k \sim q(\cdot \mid \mathcal{M}_k)$ and $u_{k\to k'} \sim g_{k\to k'}(\cdot)$, and set $(\theta_{k'}, u_{k'\to k}) = t_{k\to k'}(\theta_k, u_{k\to k'})$.

ii) Accept $(\mathcal{M}_{k'}, \theta_{k'})$ w.p. $\min\{1, \alpha_G(x_k, x_{k'})\}$ where:

$$\alpha_G(x_k, x_{k'}) := \frac{p(y \mid \theta_{k'}, \mathcal{M}_{k'})\pi(\theta_{k'})\pi(\mathcal{M}_{k'})g_{k'\to k}\left(u_{k'\to k}\right) q(\mathcal{M}_k \mid \mathcal{M}_{k'})}{p(y \mid \theta_k, \mathcal{M}_k)\pi(\theta_k)\pi(\mathcal{M}_k)g_{k\to k'}\left(u_{k\to k'}\right) q(\mathcal{M}_{k'} \mid \mathcal{M}_k)} J_{k\to k'}(\theta_k, u_{k\to k'}).$$

iii) Repeat step 1 and 2 sufficient times.

The apparent implementational problem of Algorithm 2 is the choices of $t_{k\to k'}$ and $g_{k\to k'}$ which are not straightforward, and depend on the problem being considered. Generally speaking, bad choices of $t_{k\to k'}$ and $g_{k\to k'}$ will end up with poor a performance of the algorithm. Some suggestions can be found in Brooks et al. (2003), and Dellaportas et al. (2006) study applications of the algorithm to diffusion-type models, for instance.

2.4 Pseudo-marginal MCMC

Suppose that a target $\pi \in \mathcal{P}(E)$ has a density $\pi(x)$ w.r.t. some reference measure a, say dx. In addition, suppose that $\pi(x)$ now *cannot be evaluated point-wise*. To facilitate the discussion, consider a posterior distribution given some data y and admits a density w.r.t. $d\theta$, which is also denoted by π. That is, we have:

$$\pi(\theta \mid y) = \frac{p(y \mid \theta)p(\theta)}{\int p(y \mid \theta)p(\theta)d\theta}, \tag{2.27}$$

where $\theta \mapsto p(y \mid \theta)$ is the likelihood of the observations and we assign a prior for of density θ w.r.t. $d\theta$. In many practical cases, the likelihood $p(y \mid \theta)$ cannot be evaluated point-wise, so that the Metropolis-Hastings algorithm (Algorithm 1) cannot be directly used anymore in this setting. This problem routinely appears especially in the context of latent variable models. For instance, let x be a some latent variable in the sense that one cannot directly observe from data, defined on say $(X, \mathcal{X})$.

It is well-known that joint MCMC exploring (x, θ) will suffer from slow mixing due to difficulty in updating the latent process, x, or due to strong correlation between (x, θ). This leads us to calculate the marginal likelihood $p(y \mid \theta) = \int_X p(y, x \mid \theta) dx = \int_X p(x \mid \theta) p(y \mid x, \theta) dx$. If this is the case, the ideal acceptance probability of the *marginal algorithm* is $\min\left\{1, \frac{p(y|\theta')p(\theta')q(\theta|\theta')}{p(y|\theta)p(\theta)q(\theta'|\theta)}\right\}$, which is again might not be obtained in practice due to $p(y \mid \theta)$.

To overcome this problem, we first introduce an auxiliary random variable w on a measurable space say $(W, \mathcal{W})$. The main idea of *pseudo-marginal MCMC (Beaumont, 2003; Andrieu and Roberts, 2009)* is that, whilst one cannot work with the target $\pi(dx)$, one can construct an extended target $\pi(dx, dw)$ on the product space $(E \times W, \mathcal{E} \times \mathcal{W})$ to approximate $\pi(dx)$, or to approximate the ideal acceptance probability. To do so, define a distribution $\tilde{\pi}_N$ on the product space $(E \times W, \mathcal{E} \times \mathcal{W})$:

$$\tilde{\pi}_N(dx, dw) := \pi(dx)\pi_x(dw), \tag{2.28}$$

with $\pi_x(dw) := Q_{x,N}(dw)w$ where $\{Q_{x,N}(dw)\}_{x,N \in E,\mathbb{N}}$ is a family of distributions on $(W, \mathcal{W})$ such that for each x, N:

$$\mathbb{E}\left[W_{x,N}\right] = 1, \tag{2.29}$$

where $W_{x,N} \sim Q_{x,N}(\cdot)$. We assume that $W_{x,N}$ is strictly positive w.p.1. $\{W_{x,N}\}_{x,N}$ are often referred as the *weights*. Let $f \in \mathcal{B}_b(E)$. Then, from the condition in (2.29), we have that:

$$\int f(x) w \pi(dx) Q_{x,N}(dw) = \int f(x) \pi(dx),$$

and thus *exactness* follows in this sense. Next define a proposal kernel $\tilde{Q}_N$ as follows:

$$\tilde{Q}_N\left((x, w), (dy, du)\right) := Q(x, dy) Q_{y,N}(du), \tag{2.30}$$

where $Q : E \times \mathcal{E} \to [0, 1]$. Then from (2.8), the acceptance probability is given by:

$$\alpha_{pseudo}\left((x, w), (y, u)\right) := \left\{1, \frac{\pi(y) u q(y \mid x)}{\pi(x) w q(x \mid y)}\right\}. \tag{2.31}$$

From Theorem 9, it is clear to see that the Metropolis–Hastings kernel (2.2) with (2.30) and (2.31) admits the extended target $\tilde{\pi}_N(dx, dw)$ as an invariant distribution.

Algorithm 3 Pseudo-marginal MCMC (Beaumont, 2003; Andrieu and Roberts, 2009).

Given the current variables (x, w):

i) Propose a move y via $q(\cdot \mid x)$.

ii) Given y, propose a move u via $Q_{y,N}(\cdot)$ and hence obtain $\tilde{\pi}_N(dy, du)$.

iii) Set $(x, w) = (y, u)$ w.p. (2.31).

iv) Repeat from step 1 to 3 sufficiently enough times.

Example 11. As an example, consider (2.27) of the following latent variable model. Let $\{X_n\}_{n\geq 1}$ be $\mathbb{X}-$valued *i.i.d.* latent variables, which have the density such that $X_n \overset{i.i.d.}{\sim} f_\theta(\cdot)$. Suppose that one can observe $\{X_n\}_{n\geq 1}$ only in the sense that $Y_n \mid X_n \sim g_\theta(\cdot \mid X_n)$. Then the likelihood function is given by $p(y_{1:n} \mid \theta) := \prod_{k=1}^n \int_{\mathbb{X}} f_\theta(x_k) g_\theta(y_k \mid x_k) dx_k$, which is analytically intractable. Then consider unbiased estimation of $p(y_{1:n} \mid \theta)$ via importance sampling. To do so, for $k = 1, \ldots, n$ and $i = 1, \ldots, N$, we introduce weights $w_\theta^k(u_{k,i})$ such that:

$$w_\theta^k(u_{k,i}) := \frac{f_\theta(x_{k,i}) g_\theta(y_k \mid x_{k,i})}{q_\theta(x_{k,i} \mid y_k)},$$

where $q_\theta(\cdot \mid \cdot)$ is an importance density, and here we assumed that $X_{k,i}$ can be independently sampled from $\gamma_k(\theta, u_{k,i})$ with $u_{k,i} \overset{i.i.d.}{\sim} q_u(\cdot)$ and $\gamma_k : \Theta \times \mathbb{X} \to \mathbb{X}$. Then, we might approximate the likelihood without bias as follows:

$$p(y_k \mid \theta, u_k) := \frac{1}{N} \sum_{i=1}^N w_\theta^k(u_{k,i}), \; p(y_{1:n} \mid \theta, u) := \prod_{k=1}^n p(y_k \mid \theta, u_k),$$

so that now Algorithm 3 can be applied.

To see efficiency of Algorithm 3, consider the expected acceptance rate, that is $\mathbb{E}\left[\alpha_{pseudo}\left((x,w),(y,u)\right)\right] = \int w \alpha_{pseudo}\left((x,w),(y,u)\right) Q_{x,N}(dw) Q_{y,N}(du)$. Recall that $x \mapsto \min\{1, x\}$ is a concave function. Thus we have that:

$$\begin{aligned}
\int w \alpha_{pseudo}\left((x,w),(y,u)\right) Q_{x,N}(dw) Q_{y,N}(du) &= \mathbb{E}_{Q_{x,N}(dw)}\left[\mathbb{E}_{Q_{y,N}(du)}\left[\min\left\{w, \frac{u\pi(y)q(y \mid x)}{\pi(x)q(x \mid y)}\right\}\right]\right], \\
&\leq \mathbb{E}_{Q_{x,N}(dw)}\left[\min\left\{\mathbb{E}_{Q_{y,N}(du)}\left[w\right], \mathbb{E}_{Q_{y,N}(du)}\left[\frac{u\pi(y)q(y \mid x)}{\pi(x)q(x \mid y)}\right]\right\}\right], \\
&= \mathbb{E}_{Q_{x,N}(dw)}\left[\min\left\{w, \frac{\pi(y)q(y \mid x)}{\pi(x)q(x \mid y)}\right\}\right], \\
&\leq \min\left\{\mathbb{E}_{Q_{x,N}(dw)}\left[w\right], \mathbb{E}_{Q_{x,N}(dw)}\left[\frac{\pi(y)q(y \mid x)}{\pi(x)q(x \mid y)}\right]\right\}, \\
&= \left\{1, \frac{\pi(y)q(y \mid x)}{\pi(x)q(x \mid y)}\right\},
\end{aligned}$$

therefore, *the acceptance rate of a pseudo-marginal algorithm is never greater than that of the exact marginal algorithm.* Indeed, Andrieu and Vihola (2015) show the the asymptotic variance of the exact marginal algorithm is always less than that of *a* pseudo-marginal algorithm, see Andrieu and Vihola (2015, Theorem 7).

2.5 Metropolis-adjusted Langevin algorithm

Let $\pi \in \mathcal{P}(E)$ be a target, and suppose that one wants to find a diffusion process X_t which π leaves invariant. Then, as we studied, the stationary Fokker-Planck equation has to satisfy:

$$-\sum_i \frac{\partial}{\partial x_i} b_i(x)\pi(x) + \frac{1}{2} \sum_{i,j} \frac{\partial^2}{\partial x_i \partial x_j} \Sigma_{i,j}(x)\pi(x) = 0, \tag{2.32}$$

where $\Sigma := \sigma^\top \sigma$. In other words, now we consider the inverse problem such that, given the target π, we want to solve (2.32). Then imposing the detailed balance condition gives rise to:

$$b(x)\pi(x) + \frac{1}{2}\sum_{i,j} \frac{\partial^2}{\partial x_i \partial x_j} \Sigma_{i,j}(x)\pi(x) = 0. \tag{2.33}$$

(2.33) implies that, in order to obtain samples from the target π, we may focus on reversible diffusion processes. If one sets $\Sigma = I$, then $b = \frac{1}{2}\pi^{-1}\nabla\pi = \frac{1}{2}\nabla \log \pi$. This leads to the *Langevin dynamics*:

$$dX_t = \frac{1}{2}\nabla \log \pi(X_t)dt + dW_t, \tag{2.34}$$

here we note that (2.34) is often called *Smoluchowski dynamics* as well in the context of physics. Since $-\log \pi(x)$ satisfies the Poincaré inequality, (2.34) might converge exponentially fast to the target π unless π is multimodal, see Roberts and Tweedie (1996b, Theorem 2.3, Theoren 2.4) for details.

In practice, one has to discretise (2.34) with a step size $\epsilon > 0$. Applying the Euler–Maruyama method on the interval $[0, T]$, for instance, then yields:

$$X_{i+1} = X_i + \frac{\epsilon}{2}\nabla \log \pi(X_i) + \sqrt{\epsilon}\xi_i, \tag{2.35}$$

where $\xi_i \overset{i.i.d.}{\sim} \mathcal{N}(0, I)$ and $\epsilon := \frac{T}{M}$ with a positive integer M. Although (2.34) has the desirable properties, such discretisation may destroy these properties due to the discretisation error. To offset the error induced by the discretisation, following Besag (1994), Roberts and Tweedie (1996a) use the Metropolis–Hastings kernel with the proposal:

$$Q(x, :) = \mathcal{N}\left(x + \frac{\epsilon}{2}\nabla \log \pi(x), \epsilon I\right), \tag{2.36}$$

which leads to the *Metropolis-adjusted Langevin algorithm* (MALA) as follows.

Algorithm 4 Metropolis-adjusted Langevin algorithm (Roberts and Tweedie, 1996a; Besag, 1994).

i) Given x, propose a move y via (2.36).

ii) Set $x = y$ w.p. $\alpha_{MH}(x, y)$ defined in (2.8).

iii) Repeat step 1 and 2 sufficient times.

In the same manner as Theorem 13, Roberts and Rosenthal (1998) show that, as the dimensions $d \uparrow \infty$, the optimal acceptance rate for the algorithm is 0.574 for Algorithm 4. Critically, in the case of MALA, one needs to scale the parameter ϵ as $\epsilon = \frac{\ell^2}{d^{1/3}}$. It turns out the computational complexity, as $d \uparrow \infty$, of MALA is $\mathcal{O}(d^{1/3})$ compared with one of the random walk Metropolis-Hastings is $\mathcal{O}(d)$. Therefore, MALA is much more efficient than the random walk Metropolis-Hastings. This is not a surprising result since if one were able to obtain samples directly from (2.34), then the acceptance probability would be close to 1 by the construction.

2.6 Hamiltonian Monte Carlo

Let $q \in \mathbb{R}^d$ be the variables of interest. Hamiltonian Monte Carlo (HMC) first introduces auxiliary variables $p \in \mathbb{R}^d$. In the context of molecular dynamics, q and p are often called *position* variables and *momentum* variables respectively. Let $D := 2d$. The total energy of the system is given by the (separable) *Hamiltonian* $H : \mathbb{R}^D \mapsto \mathbb{R}$ such that:

$$H(q, p) := U(q) + K(p), \tag{2.37}$$

where $U : \mathbb{R}^d \mapsto \mathbb{R}$ is called the *potential function* and $K : \mathbb{R}^d \mapsto \mathbb{R}$ is called the *kinetic function*. Also, in this setting, $\mathbb{R}^D$ is called the *phase space*. In statistics, $U(q)$ might be proportional to the (log) negative target distribution from which one wants to sample. Also, the kinetic function is often of the form:

$$K(p) = \frac{1}{2} p^\top M^{-1} p, \tag{2.38}$$

where the $d \times d$ matrix M is called the *mass matrix*. This class of kinetic functions is often called Gaussian kinetic function (Betancourt, 2017). We note that although $K(\cdot)$ can depend on q, we will restrict ourselves to a separable Hamiltonian. Let $t \in \mathbb{R}_+$ be an auxiliary time index. Then the *Hamiltonian dynamics* is given by:

$$\begin{cases} \frac{dq(t)}{dt} = \nabla_p H(q(t), p(t)) = M^{-1} p(t), \\ \frac{dp(t)}{dt} = -\nabla_q H(q(t), p(t)) = -\nabla U(q(t)). \end{cases} \tag{2.39}$$

Let $x := (q, p) \in \mathbb{R}^D$ and define the matrix:

$$J := \begin{bmatrix} 0_{d\times d} & I_{d\times d} \\ -I_{d\times d} & 0_{d\times d} \end{bmatrix}. \tag{2.40}$$

Using this, (2.39) can be equivalently formulated as:

$$\frac{dx(t)}{dt} = J \nabla H(x). \tag{2.41}$$

Given an initial condition $\{q(0), p(0)\} = (p_0, q_0) =: x_0 \in \mathbb{R}^D$, if $H(x)$ is bounded below and $\nabla H(x)$ is locally Lipschitz continuous, then existence and uniqueness of (2.41) follows, see e.g., Stoltz and Rousset (2010). Then, (2.41) introduces the *flow* $\{\Phi_t\}$ in the sense that $\Phi_t(x_0)$ is the value at time t of the solution $x(t)$ of the Hamiltonian dynamics (2.41) with an initial condition x_0.

Next we study some properties of flow $\{\Phi_t\}$. A mapping $\Phi_t : \mathbb{R}^D \mapsto \mathbb{R}^D$ is said to be *symplectic* if for any point $x \in \mathbb{R}^D$:

$$\nabla \Phi^\top J \nabla \Phi^\top = J, \tag{2.42}$$

holds, where J is defined in (2.40). Notice that the matrix J satisfies $J^\top = -J = J^{-1}$. Then we have the following propositions.

Proposition 14. *The flow* $\{\Phi_t\}$ *induced by* (2.41) *is symplectic.*

Proof. See, e.g., Neal (2011); Betancourt (2017). □

As a consequence of Φ_t being symplectic, we immediately have the following important result.

Proposition 15. *The determinant of* $\nabla\Phi$ *is always* 1 *for any point* $x \in \mathbb{R}^D$.

Proof. Assume that $\det(J) \neq 0$. Then from Proposition 14, for any point $x \in \mathbb{R}^D$, we have that:

$$\det\left(\nabla\Phi^\top\right)\det(J)\det(\nabla\Phi) = \det(J),$$

and this gives rise to $\det(\nabla\Phi) = 1$ since $\det(J) \neq 0$ and thus $\det\left(\nabla\Phi^\top\right)\det(\nabla\Phi) = 1$. □

Essentially, what Proposition 15 says is following. Let Leb^D denote the Lebesgue measure on $\mathbb{R}^D$. Then for any Borel $A \in \mathcal{B}(\mathbb{R}^D)$, we have $\text{Leb}^D(\Phi_t(A)) = \text{Leb}^D(A)$ for any $t \in \mathbb{R}_+$. Equivalently, for bounded $f \in \mathcal{B}_b(\mathbb{R}^D)$, we have that $\int_{\Phi_t(A)} f(q,p)\mu(dq)\nu(dp) = \int_A f(\Phi_t(A))\det(\nabla\Phi_t(A))\,\mu(dq)\nu(dp) = \int_A f(\Phi_t(A))\mu(dq)\nu(dp)$ for any $t \in \mathbb{R}_+$ and Lebesgue measures. Proposition 15 is called *volume preservation*. Also we have the following.

Proposition 16. *For any* $t \in \mathbb{R}_+$, *the flow* $\{\Phi_t\}$ *induced by* (2.41) *and the Hamiltonian function* $H()$ *in* (2.37) *satisfy* $H \circ \Phi_t = H$.

Proof. $\frac{dH(x(t))}{dt} = \frac{\partial H(x)}{\partial q}\frac{dq(t)}{dt} + \frac{\partial H(x)}{\partial p}\frac{dp(t)}{dt} = \frac{\partial H(x)}{\partial q}\frac{\partial H(x)}{\partial p} - \frac{\partial H(x)}{\partial p}\frac{\partial H(x)}{\partial q} = 0$, thus $H(\Phi_t(x_0)) = H(x_0)$. □

Proposition 16 is called *energy conservation*. From *volume preservation* (Proposition 15) and *energy conservation* (Proposition 16), the following key claim immediately follows.

Theorem 14. *For each* t, *the probability measure in* $\mathbb{R}^D$ *the density* $Z^{-1}exp(-H(q,p))$ *w.r.t. Lebesgue is preserved by the flow* Φ_t *induced by* (2.41)*:*

$$\int_{\Phi_t(A)} Z^{-1}exp(-H(q,p))\,\mu(dq)\nu(dp) = \int_A Z^{-1}exp(-H(q,p))\,\mu(dq)\nu(dp),$$

for any Borel $A \in \mathcal{B}(\mathbb{R}^D)$, *where* Z *is the normalising constant.*

Proof. For any Borel $A \in \mathcal{B}(\mathbb{R}^D)$:

$$\begin{aligned}\int_{\Phi_t(A)} Z^{-1}\exp(-H(q,p))\,\mu(dq)\nu(dp) &= \int_A Z^{-1}\exp(-H(\Phi_t(A)))\mid\det(\Phi_t(A))\mid\mu(dq)\nu(dp),\\ &= \int_A Z^{-1}\exp(-H(A))\,\mu(dq)\nu(dp).\end{aligned}$$

□

The key implication of Theorem 14 is as follows. Assume that now the potential function $U(q)$ is the negative logarithm of the density of the target, say $\pi(dq)$. That is, we have that:

$$\pi(dq) = Z_q^{-1}\exp(-U(q))\,\mu(dq),\ Z_q = \int_{\mathbb{R}^d}\exp(-U(q))\,\mu(dq).$$

This means that $\exp(-H(q,p))$ has the unnormalised density $\exp(-U(q)) \times \exp\left(-\frac{1}{2}p^\top M^{-1}p\right)$, since q and p are statistically independent. Clearly, the target distribution $\pi(dq)$ is the $q-$marginal of $\exp(-H(q,p))$, and $p-$marginal of it is the standard Gaussian distribution. Then, in this set-up, Theorem 14 implies that the following ideal algorithm will construct a Markov process in $\mathbb{R}^d$ reversible w.r.t. the target $\pi(dq)$, see e.g., Sanz-Serna (2014, Section 9) or Bou-Rabee and Sanz-Serna (2018, Section 5).

Proposition 17. *Let $L > 0$ denote the duration parameter. Define the transitions $q_n \mapsto q_{n+1}$ in $\mathbb{R}^d$ by the following procedure.*

i) Sample p_n from a $d-$dimensional standard Gaussian distribution $\mathcal{N}(0, M)$.

ii) Obtain (q_{n+1}, p_{n+1}) by evolving the Hamiltonian dynamics (2.41) over the time interval $[0, L]$.

iii) Let $\mathfrak{proj}^{\mathfrak{q}}$ denote the projection on the $q-$component. Set $q_{n+1} = q(L) = \mathfrak{proj}^{\mathfrak{q}}(q_n, p_n)$ and discard p_{n+1}.

Then the Markov process $q_n \mapsto q_{n+1}$ leaves the target $\pi(dq)$ invariant marginally. Also, the Markov process is reversible w.r.t. the target $\pi(dq)$.

Proof. The first claims follows immediately from Theorem 14. The Markovian property follows from the fact that the past enters the computation of q_1 only through q_0. We will study later such reversibility by studying the flow $\{\Phi_t\}$. □

Next we study further the properties of the flow $\{\Phi_t\}$. The linear map $S : \mathbb{R}^D \to \mathbb{R}^D$ is said to be linear *involution* in $\mathbb{R}^D$ if $S(S(x)) = x$ for any $x \in \mathbb{R}^D$. In particular, we consider the *momentum flip* involution:

$$S(q,p) = S(q-p). \tag{2.43}$$

Clearly, the momentum flip involution is linear involution. Assume that the Hamiltonian $H(q,p)$ is a even function of the momentum, i.e., $H(q,p) = H(q,-p)$. Equivalently, $H(q,p) = H(S(q,p))$. Suppose that $(q(t), p(t))$ is a solution of (2.39), and set $(\hat{q}(t), \hat{p}(t)) := (q(-t), -p(-t))$. Then observe that:

$$\frac{d\hat{q}(t)}{dt} = \frac{dq(t)}{d(-t)} = \nabla_p H(q(-t), p(-t)) = \nabla_p H(q(-t), -p(-t)) = \nabla_p H(\hat{q}(t), \hat{p}(t)),$$
$$\frac{d\hat{p}(t)}{dt} = \frac{d(-p(t))}{d(-t)} = -\nabla_q H(q(-t), -p(-t)) = -\nabla_q H(\hat{q}(t), \hat{p}(t)).$$

From this observation, we can obtain *reversibility* (w.r.t. S) of the flow $\{\Phi_t\}$:

$$S \circ \Phi_t = \Phi_{-t} \circ S,$$
$$\iff \Phi_{-t} = S \circ \Phi_t \circ S. \tag{2.44}$$

We note that t*he flow $\{\Phi_t\}$ itself is not reversible*, although applying the momentum flip involution S makes $\{\Phi_t\}$ reversible. Also notice that $\Phi_t \circ \Phi_{-t} = Id$, so we have that:

$$\Phi_{-t} = \Phi_t^{-1}. \tag{2.45}$$

The property (2.45) is called *symmetry*. From reversibility and symmetry, we have that:

$$\Phi_t^{-1} = S \circ \Phi_t \circ S = \Phi_{-t}.$$

The following result follows from (2.44) and (2.45), and will be useful in the sequel.

Proposition 18. *Bou-Rabee and Sanz-Serna (2018, Proposition 2.5). The flow $\{\Phi_t\}$ and the momentum flip S satisfy:*

$$|\det\left(\nabla\Phi_t\left(S(\Phi_t)\right)\right)| = |\det\left(\nabla\Phi_t\left(x\right)\right)|^{-1}.$$

All in all, we summarise the important properties of the flow $\{\Phi_t\}$ as the following theorem.

Theorem 15. *Let the flow $\{\Phi_t\}$ by (2.41), the Hamiltonian function $H(q,p)$ in (2.37) and S be the momentum flip in (2.43). Then we have the followings:*

i) $\det\left(\nabla\Phi_t\right) = 1$.

ii) $H \circ \Phi_t = H$.

iii) $S \circ \Phi_t = \Phi_{-t} \circ S$.

iv) $\Phi_{-t} = \Phi_t^{-1}$.

v) Let $\tilde{\pi}(dq,dp) \propto exp\left(-H(q,p)\right)\mu(dq)\nu(dp)$. Then any Borel $A \in \mathcal{B}(\mathbb{R}^D)$, the push forward measure $\tilde{\pi}(\Phi_t^{-1}(A))$ is equal to $\tilde{\pi}(A)$.

If one had samples from ideal algorithm (Proposition 17), then such samples would leave the target invariant as we studied. However, one cannot, in general, solve the Hamiltonian dynamics analytically (2.41). This leads us to resort to numerical integrators. Some care is needed here. As we studied, the flow $\{\Phi_t\}$ has several desirable properties, see ,e.g., Theorem 15. We do not want such properties to break down due to discretisation. Although it *cannot preserve energy conservation*, the best-known *volume preserving*, *reversible* w.r.t. S algorithm to integrate the Hamiltonian dynamics numerically is the *Verlet/leapfrog integrator*: see, e.g. Bou-Rabee and Sanz-Serna (2018, Section 4) and Stoltz and Rousset (2010, Section 1).

The leapfrog integrator is based on *Strang's splitting*. Recall that, since it is separable, one can split the Hamiltonian into two parts:

$$H^{(1)}(q) := U(q),\ H^{(2)}(p) := \frac{1}{2}p^\top M^{-1}p, \tag{2.46}$$

clearly $H(q,p) = H^{(1)}(q) + H^{(2)}(p)$. As for $H^{(1)}(q)$, the corresponding dynamics are $\frac{dq(t)}{dt} = \nabla_p H^{(1)}(q) = 0$ and $\frac{dp(t)}{dt} = -\nabla_q H^{(1)}(q) = -\nabla U(q)$. Also, as for $H^{(2)}(p)$, we have that $\frac{dq(t)}{dt} = \nabla_p H^{(2)}(p) = M^{-1}p$ and $\frac{dp(t)}{dt} = -\nabla_q H^{(2)}(p) = 0$. Then let $\epsilon \in \mathbb{R}_+$ and $L \in \mathbb{N}$. The leapfrog integrator proceeds as follows.

Given initial points $(q_0, p_0) \in \mathbb{R}^D$, one can iterate L times following step:

$$\begin{cases} p_{k+1/2} &= p_k - \frac{\epsilon}{2}\nabla U(q_k), \\ q_{k+1} &= q_k + \epsilon M^{-1} p_{k+1/2}, \\ p_{k+1} &= p_{k+1/2} - \frac{\epsilon}{2}\nabla U(q_{k+1}), \end{cases} \tag{2.47}$$

for $k \in \{0, \cdots, L-1\}$. This sequence defines a discrete dynamical system for $k \in \{0, \cdots, L-1\}$ as follows:

$$\begin{aligned}(q_{k+1}, p_{k+1}) &= \Xi^{(1)}_{\epsilon/2} \circ \Xi^{(2)}_{\epsilon} \circ \Xi^{(1)}_{\epsilon/2}(q_k, p_k) \\ &=: \Psi^{(1)}_{\epsilon}(q_k, p_k),\end{aligned}$$

where $\Xi^{(1)}_{\epsilon}$, $\Xi^{(2)}_{\epsilon} : \mathbb{R}^D \to \mathbb{R}^D$ are given for all $(q, p) \in \mathbb{R}^D$ by:

$$\Xi^{(1)}_{\epsilon}(q, p) := (q, p - \frac{\epsilon}{2}\nabla U(q)), \ \Xi^{(2)}_{\epsilon}(q, p) := (q + \epsilon M^{-1} p, p). \tag{2.48}$$

By iterating this process, we can define the sequence for $k \geq 1$:

$$\Psi^{\circ(k+1)}_{\epsilon} = \Psi^{\circ(k)}_{\epsilon} \circ \Psi^{(1)}_{\epsilon}. \tag{2.49}$$

It is clear to see that the Jacobian of $\Xi^{(1)}_{\epsilon}$, denoted by $J(\Xi^{(1)}_{\epsilon})$, is given by:

$$J(\Xi^{(1)}_{\epsilon}) = \begin{pmatrix} 1 & 0 \\ -\frac{\epsilon}{2}\nabla^2 U(q) & 1 \end{pmatrix},$$

thus we conclude that $\det\left(J(\Xi^{(1)}_{\epsilon})\right) = 1$. Also, one can show that $\det\left(J(\Xi^{(2)}_{\epsilon})\right) = 1$. As a result, we have the following. Note that the composition of two volume preserving mappings is itself volume preserving, see Bou-Rabee and Sanz-Serna (2018, Proposition 2.2).

Proposition 19. *The approximated flow* $\{\Psi^{\circ(k)}_{\epsilon}\}$ *for* $k \in \{0, \cdots, L\}$ *in* (2.49) *preserves volume preservation of the exact* flow $\{\Phi_t\}$ for any $\epsilon \in \mathbb{R}_+$.

To see reversibility of $\{\Psi^{\circ(k)}_{\epsilon}\}$ w.r.t. the momentum flip involution S, consider (here we set $M = I_d$ without loss of generality):

$$\begin{cases} q_{k+1} &= q_k + \epsilon p_k - \frac{\epsilon^2}{2}\nabla U(q_k), \\ p_{k+1} &= p_k - \frac{\epsilon}{2}\nabla U(q_k) - \frac{\epsilon}{2}\nabla U(q_{k+1}). \end{cases} \tag{2.50}$$

Using this expression, we have:

$$\begin{aligned}\mathfrak{proj}^{\mathfrak{q}}(\Psi^{(1)}_{\epsilon}(q_k, -p_k)) &= q_{k+1} - \epsilon p_{k+1} - \frac{\epsilon^2}{2}\nabla U(q_{k+1}), \\ &= q_k + \epsilon p_k - \frac{\epsilon^2}{2}\nabla U(q_k) - \epsilon\left(p_k - \frac{\epsilon}{2}\nabla U(q_k) - \frac{\epsilon}{2}\nabla U(q_{k+1})\right) - \frac{\epsilon^2}{2}\nabla U(q_{k+1}), \\ &= q_k,\end{aligned}$$

and:

$$\begin{aligned}\mathfrak{proj}^{\mathfrak{p}}(\Psi_{\epsilon}^{(1)}(q_k,-p_k)) &= -\left(-p_{k+1}-\frac{\epsilon}{2}\nabla U(q_{k+1})-\frac{\epsilon}{2}\nabla U(\mathfrak{proj}^{\mathfrak{q}}(\Psi_{\epsilon}^{(1)}(q_k,-p_k)))\right),\\ &= -\left(-\left(p_k-\frac{\epsilon}{2}\nabla U(q_k)-\frac{\epsilon}{2}\nabla U(q_{k+1})\right)-\frac{\epsilon}{2}\nabla U(q_{k+1})-\frac{\epsilon}{2}\nabla U(q_k)\right),\\ &= p_k.\end{aligned}$$

These observations imply $S\circ\Psi_{\epsilon}^{(1)}\circ S=(q_k,-p_k)=\left(\Psi_{-\epsilon}^{(1)}\right)$ since:

$$\begin{aligned}\mathfrak{proj}^{\mathfrak{q}}(\Psi_{-\epsilon}^{(1)}(q_k,p_k)) &= q_k,\\ \mathfrak{proj}^{\mathfrak{p}}(\Psi_{-\epsilon}^{(1)}(q_k,p_k)) &= -p_k.\end{aligned}$$

holds. Thus we have the following.

Proposition 20. *The approximated flow* $\{\Psi_{\epsilon}^{\circ(k)}\}$ *for* $k\in\{0,\cdots,L\}$ *(and* $\Psi_{\epsilon}^{(1)}$*) in* (2.49) *preserves* reversibility *(w.r.t. S) of the flow* $\{\Phi_t\}$.

As we mentioned, the approximated flow $\{\Psi_{\epsilon}^{\circ(k)}\}$ does not preserve energy conservation, and a natural strategy for correcting this bias will be to use the Metropolis- Hastings scheme. This is pretty much the same as MALA. After L leapfrog steps, one has the last points of the numerical trajectory (q_L,p_L) with transition:

$$\mathbb{Q}(q',p'\mid q_0,p_0)=\delta(q'-q_L)\delta(p'-p_L).$$

Consider the ratio of the transitions w*ithout the momentum flip:*

$$\frac{\mathbb{Q}(q_0,p_0\mid q_L,p_L)}{\mathbb{Q}(q_L,p_L\mid q_0,p_0)}=\frac{0}{1},$$

so the Metropolis- Hastings acceptance probability will be always zero. This is because, without the momentum flip, the approximated flow $\{\Psi_{\epsilon}^{\circ(k)}\}$ proceeds only forwards (Betancourt, 2017). Again, the approximated flow $\{\Psi_{\epsilon}^{\circ(k)}\}$ itself is not reversible. If we modify (but not necessarily) the bias via the Metropolis- Hastings scheme, then we need to apply the momentum flip to the the approximated flow $\{\Psi_{\epsilon}^{\circ(k)}\}$, which gives rise to the reversible proposal:

$$\mathbb{Q}(q',p'\mid q_0,p_0)=\delta_{q_L}(q'-q_L)\delta(p'+p_L).$$

$\tilde{\pi}(q,p)\propto exp\,(-H(q,p))$. In this case, the Metropolis- Hastings acceptance probability becomes:

$$\begin{aligned}\alpha_H(q_L,-p_L) &:= \min\left(1,\frac{\mathbb{Q}(q_0,p_0\mid q_L,-p_L)\tilde{\pi}(q_L,-p_L)}{\mathbb{Q}(q_L,-p_L\mid q_0,p_0)\tilde{\pi}(q_0,p_0)}\right),\\ &= \min\left(1,\frac{\delta(q_0-q_0)\delta(-p_0+p_0)\tilde{\pi}(q_L,-p_L)}{\delta(q_L-q_L)\delta(-p_L+p_L)\tilde{\pi}(q_0,p_0)}\right),\\ &= \min\left(1,\frac{\exp\left(-H(q_L,-p_L)\right)}{\exp\left(-H(q_0,p_0)\right)}\right),\end{aligned}\tag{2.51}$$

as desired. Now we are ready to introduce the Hamiltonian Monte Carlo (HMC), also known as the hybrid Monte Carlo, presented first in Duane et al. (1987). Notice that if the exact flow $\{\Phi_t\}$ were available, that is $\Phi_L = \Psi_\epsilon^{\circ(L)}$, then:

$$\frac{\exp\left(-H(q_L, -p_L)\right)}{\exp\left(-H(q_0, p_0)\right)} = 1,$$

due to Proposition 16. Therefore, the better numerical scheme $\Psi_\epsilon^{\circ(L)}$ can end up with the higher probability of acceptance.

Algorithm 5 Hamiltonian Monte Carlo (HMC) (Duane et al., 1987).

i) Draw p_n from $\mathcal{N}(0, M)$ given q_n.

ii) Obtain $(q_{n+1}^*, p_{n+1}^*) = \Psi_\epsilon^{\circ(L)}(q_n, p_n)$ by iterating (2.49) $L-1$ times.

iii) Set $q_{n+1} = q_{n+1}^*$ w.p.

$$\alpha_H(q_{n+1}^*, -p_{n+1}^*) = \min\left(1, \frac{\exp\left(-H(q_{n+1}^*, -p_{n+1}^*)\right)}{\exp\left(-H(q_n, p_n)\right)}\right),$$

otherwise set $q_{n+1} = q_n$.

iv) Discard p_{n+1}^*.

v) Repeat from the first step to the fourth step sufficiently large n times.

Following Stoltz and Rousset (2010, Section 2) and Sanz-Serna (2014) closely, we study the validity of Algorithm 5.

Lemma 4. *Let μ be a Lebesgue measure on $\mathbb{R}^D$ which is preserved by the momentum flip S, that is for any $A \in \mathcal{B}(\mathbb{R}^D)$, $\mu(A) = \mu(S(A)) = \mu(S(A)^{-1})$. Let K be a kernel on $\mathbb{R}^D$ such that:*

$$\int_A \mu(dx) K(x, B) = \int_B \mu(dy) K(S(y), S(A)), \tag{2.52}$$

for any $A, B \in \mathcal{B}(\mathbb{R}^D)$. Then we have that:

i) The measure μ is invariant w.r.t. K.

ii) At stationarity, the Markov chain $\{\xi_i\}$ generated by K is statistically same as the chain $\{S(\xi_i)\}$.

Proof. Take $A = \mathbb{R}^D$, then we have that:

$$\begin{aligned}\int_{\mathbb{R}^D} \mu(dx) K(x, B) &= \int_B \mu(dy) K(S(y), \mathbb{R}^D), \\ &= \int_B \mu(dy) = \mu(B),\end{aligned}$$

this proves that the measure μ is invariant w.r.t. K. Next, by the definition, we obtain:

$$\begin{aligned}\mathbb{P}\left(S(\xi_n) \in S(A) \mid S(\xi_{n+1}) \in S(B)\right) &= \mathbb{P}\left(\xi_n \in A \mid \xi_{n+1} \in B\right), \\ &= \frac{\mathbb{P}\left(\xi_n \in A, \xi_{n+1} \in B\right)}{\mathbb{P}\left(\xi_{n+1} \in B\right)}.\end{aligned}$$

Recall that, for any n, $\mathbb{P}\left(\xi_{n+1} \in B\right) = \mathbb{P}\left(\xi_n \in B\right)$. Set $y = S(x)$ and using $\mu(S(dx)) = \mu(dx)$ give rise to:

$$\begin{aligned}\int_A \mu(dx)K(x,B) &= \int_{S(B)} \mu(S(dx))K(x,S(A)), \\ &= \int_{S(B)} \mu(dx)K(x,S(A)),\end{aligned}$$

and this implies $\mathbb{P}\left(\xi_n \in A, \xi_{n+1} \in B\right) = \mathbb{P}\left(\xi_n \in S(B), \xi_{n+1} \in S(A)\right)$. As a result, we have that:

$$\begin{aligned}\mathbb{P}\left(S(\xi_n) \in S(A) \mid S(\xi_{n+1}) \in S(B)\right) &= \frac{\mathbb{P}\left(\xi_n \in A, \xi_{n+1} \in B\right)}{\mathbb{P}\left(\xi_{n+1} \in B\right)}, \\ &= \frac{\mathbb{P}\left(\xi_n \in S(B), \xi_{n+1} \in S(A)\right)}{\mathbb{P}\left(\xi_n \in S(B)\right)}, \\ &\quad \mathbb{P}\left(\xi_{n+1} \in S(A) \mid \xi_n \in S(B)\right).\end{aligned}$$

□

Lemma 5. *Let μ be a Lebesgue measure on $\mathbb{R}^D$ which is preserved the momentum flip S, that is for any $A \in \mathcal{B}(\mathbb{R}^D)$, $\mu(A) = \mu(S(A)) = \mu(S(A)^{-1})$. Assume that K is a Markov kernel in $\mathbb{R}^D$ such that the two measures:*

$$K(S(y), S(dx))\mu(du),\ K(x,dy)\mu(dx)$$

are equivalent on $\mathbb{R}^D \times \mathbb{R}^D$ so that one can define a function $r(x,y)$ such that:

$$r(x,y) := \frac{K(S(y), S(dx))\mu(dy)}{K(x,dy)\mu(dx)}. \tag{2.53}$$

Define $\xi_n \mapsto \xi_{n+1} \in \mathbb{R}^D$ by:

i) Draw ξ^*_{n+1} from $K(\xi_n, \cdot)$.

ii) Set $\xi_{n+1} = \xi^*_{n+1}$ w.p. $\min(1, r(\xi_n, \xi^*_{n+1}))$ otherwise set $\xi_{n+1} = S(\xi_n)$.

Then the induced Markov chain satisfies (2.52) and thus μ is an invariant measure.

Proof. The Kernel Q of this chain is:

$$\begin{aligned}Q(x,dy) &= (1 \wedge r(x,y))K(x,dy) + (1 - \alpha(x))\delta_{S(x)}(dy), \\ \alpha(x) &:= \int (1 \wedge r(x,y))K(x,dy).\end{aligned}$$

We want to show that:

$$\underbrace{(1 \wedge r(x,y))K(x,dy)\mu(dx)}_{A} + \underbrace{(1-\alpha(x))\delta_{S(x)}(dy)\mu(dx)}_{B} =$$
$$\underbrace{(1 \wedge r(S(y),S(x)))K(S(y),S(dx))\mu(dy)}_{A} + \underbrace{(1-\alpha(S(y)))\delta_{y}(S(dx))\mu(dy)}_{B},$$

First, notice that:

$$\begin{aligned}\frac{1}{r(S(y),S(x))} &= \frac{K(S(y),S(dx))\mu(S(dy))}{K(S(S(x)),S(S(dy)))\mu(S(dx))} \\ &= \frac{K(S(y),S(dx))\mu(dy)}{K(x,dy)\mu(dx)} = r(x,y),\end{aligned}$$

since $\mu(dx) = \mu(S(dx))$ and $S \circ S(x) = x$. Then, using the fact $\min(1,r) = r\min\left(1,\frac{1}{r}\right)$ gives rise to:

$$\min\left(1, r(x,y)\right) K(x,dy)\mu(dx) = \min\left(1, r(S(y),S(x))\right) r(x,y) K(x,dy)\mu(dx),$$

and by the definition of $r(x,y)$ we have that:

$$\min\left(1, r(S(y),S(x))\right) r(x,y) K(x,dy)\mu(dx) = \min\left(1, r(S(y),S(x))\right) K(S(y),S(dx))\mu(dy),$$

thus the A terms are equal. Next we compare the B terms. Let $f \in \mathcal{B}_b(\mathbb{R}^D \times \mathbb{R}^D)$ and consider the change of variables $x = S(x')$. Then by we have that:

$$\begin{aligned}\int_{\mathbb{R}^D\times\mathbb{R}^D} f(x,y)(1-\alpha(S(y)))\delta_y(S(dx))\mu(dy) &= \int_{\mathbb{R}^D\times\mathbb{R}^D} f(S(x'),y)(1-\alpha(S(y)))\delta_y(dx')\mu(dy), \\ &= \int_{\mathbb{R}^D} f(S(y),y)(1-\alpha(S(y)))\mu(dy).\end{aligned}$$

Next take $y = S(x)$. Then we obtain:

$$\begin{aligned}\int_{\mathbb{R}^D} f(S(y),y)(1-\alpha(S(y)))\mu(dy) &= \int_{\mathbb{R}^D} f(S(S(x)),S(x))(1-\alpha(S(S(x))))\mu(S(dx)), \\ &= \int_{\mathbb{R}^D} f(x,S(x))(1-\alpha(x))\mu(dx), \\ &= \int_{\mathbb{R}^D\times\mathbb{R}^D} f(x,y)(1-\alpha(x))\delta_{S(x)}(dy)\mu(dx),\end{aligned}$$

thus the B terms, and thus the claim follows from Lemma 4. □

Lemma 6. *Let μ be the Lebesgue measure on $\mathbb{R}^D$ with the density which is proportional to $\exp(-H())$ w.r.t. dy with $H \circ S = S$. Besides, assume that the numerical flow $\{\Psi_\epsilon^{\circ(k)}\}$ is symmetric w.r.t. the momentum flip S and volume preserving. Define a transition kernel by:*

$$K(x,dy) = \delta_{\Psi_\epsilon^{\circ(L)}(x)}dy. \tag{2.54}$$

Then such μ and K satisfy conditions in Lemma 5 and the Metropolis-Hastings ratio $r(x,y)$ in (2.53) is given by $\frac{\exp\left(-H(\Psi_\epsilon^{\circ(L)}(x)\right)}{\exp(-H(x))}$.

Proof. Let $f \in \mathcal{B}_b(\mathbb{R}^D \times \mathbb{R}^D)$. We first consider the numerator in (2.53), and define:

$$\begin{aligned} I_N &:= \int_{\mathbb{R}^D \times \mathbb{R}^D} f(x,y) K(S(y), S(dx)) \mu(dy) \\ &= \int_{\mathbb{R}^D \times \mathbb{R}^D} f(x,y) \delta_{\Psi_\epsilon^{\circ(L)}(S(y))} S(dx) \mu(dy). \end{aligned}$$

Define $x = S(x')$ and this gives, recall that $S \circ \Psi_\epsilon^{(L)} = \left(\Psi_\epsilon^{(L)}\right)^{-1} \circ S$,

$$\begin{aligned} I_N &= \int_{\mathbb{R}^D \times \mathbb{R}^D} f(S(x'), y) \delta_{\Psi_\epsilon^{\circ(L)}(S(y))} S(S(dx')) \mu(dy) \\ &= \int_{\mathbb{R}^D} f\left(S(\Psi_\epsilon^{\circ(L)}(S(y))), y\right) \mu(dy), \\ &= \int_{\mathbb{R}^D} f\left(\left(\Psi_\epsilon^{(L)}\right)^{-1}(y), y\right) \exp(-H(y)) dy, \end{aligned}$$

and then change of the variable $x = \left(\Psi_\epsilon^{(L)}\right)^{-1}(y)$ gives:

$$\begin{aligned} I_N &= \int_{\mathbb{R}^D} f\left(x, \Psi_\epsilon^{(L)}(x)\right) \exp(-H(\Psi_\epsilon^{(L)}(x))) \mid \nabla \det\left(\left(\Psi_\epsilon^{(L)}\right)(x)\right) \mid dx, \\ &= \int_{\mathbb{R}^D} f\left(x, \Psi_\epsilon^{(L)}(x)\right) \exp(-H(\Psi_\epsilon^{(L)}(x))) dx. \end{aligned}$$

As for the denominator of (2.53), we have that:

$$\begin{aligned} I_D &:= \int_{\mathbb{R}^D \times \mathbb{R}^D} f(x,y) K(x, dy) \mu(dx), \\ &= \int_{\mathbb{R}^D \times \mathbb{R}^D} f(x,y) \delta_{\Psi_\epsilon^{\circ(L)}(x)} dy \mu(dx), \\ &= \int_{\mathbb{R}^D} f\left(x, \Psi_\epsilon^{\circ(L)}(x)\right) \exp(-H(x)) dx. \end{aligned}$$

Thus we conclude that $\delta_{\Psi_\epsilon^{\circ(L)}(S(y))} S(dx)\mu(dy)$ and $\delta_{\Psi_\epsilon^{\circ(L)}(x)} dy \mu(dx)$ are equivalent on $\mathbb{R}^D \times \mathbb{R}^D$, and taking $f = 1$ yields the claim. □

Theorem 16. *The chain induced by Algorithm 5 leaves marginally* $\exp(-U(x))$ *invariant.*

Proof. The Kernel Q of this is given by:

$$\begin{aligned} Q(x, dy) &= (1 \wedge r(x,y)) \delta_{\Psi_\epsilon^{\circ(L)}(x)} dy + (1 - \alpha(x)) \delta_{S(x)}(dy), \\ r(x,y) &= \frac{\exp(-H(\Psi_\epsilon^{(L)}(x)))}{\exp(-H(x))}, \\ \alpha(x) &:= \int (1 \wedge r(x,y)) \delta_{\Psi_\epsilon^{\circ(L)}(x)} dy. \end{aligned}$$

Lemma 6 and Lemma 5 ensure that this kernel induced by Algorithm 5 admits well-defined the Metropolis-Hastings ratio, and also that chains satisfy the condition (2.52). Thus the claim follows from Lemma 4. □

Let L be fixed and take $M = I$ in Algorithm 5. Then L steps proposals for (p, q) of the HMC can be expressed as:

$$\begin{cases} q_{t+L\epsilon} & = q_t + L\epsilon^2 \nabla \log U(q_t)/2 + \epsilon^2 \sum_{i=1}^{L-1} (L-i) \nabla \log U(q_{t+i\epsilon}) + L\epsilon p_t, \\ p_{t+L\epsilon} & = p_t + \epsilon^2 \nabla \log U(q_t)/2 + \epsilon \sum_{i=1}^{L-1} (L-i) \nabla \log U(q_{t+i\epsilon}) + \epsilon \nabla \log U(q_{t+L\epsilon})/2. \end{cases} \tag{2.55}$$

Upon observing the similarity between (2.55) and the proposal of MALA (Algorithm 4), Livingstone et al. (2019) establish, under appropriate conditions, the results which a version (fixed L) of HMC will be and not be geometrically ergodic inspired by the results such as Proposition 13 and Theorem 12. In Durmus et al. (2017), the irreducibility and geometric ergodicity of HMC with either fixed or random number of steps of the symplastic integrator is studied, under more analytical assumptions.

Optimal scaling is studied in Beskos et al. (2013b). This paper shows that for *i.i.d.* targets, the leapfrog step size ϵ has to be scaled as $\epsilon = \frac{\ell}{d^4}$ and it turns out that HMC requires $\mathcal{O}(d^{1/4})$ steps to explore the state space as $d \uparrow \infty$. Also, they identify the asymptotically optimal acceptance probability analytically is 0.651.

2.7 Advanced Hamiltonian Monte Carlo

As we noted in subsection 2.6, one has to decrease the leapfrog step size ϵ with order $\mathcal{O}(d^{-1/4})$, or the acceptance probability will degenerate into 0 at the end. To address this problem, Beskos et al. (2011, 2013a) study the HMC on Hilbert spaces. The aim of this section is to present well-defined HMC for targets in (2.56) developed in Beskos et al. (2011, 2013a).

To facilitate argument, suppose that we are interested in estimating path of (scalar) diffusion process in a certain interval $[0, l]$ where $l > 0$. In this case, the distributions of interest are defined on the infinite-dimensional Hilbert space such that $\mathcal{H} = L^2([0, l], \mathbb{R})$. We want to sample from a target distribution which is obtained as a *change of Gaussian measures on a certain (separable) Hilbert space* $\mathcal{H}$. To be precise, let Π_0 be a centred Gaussian law denoted by $\mathcal{N}(0, \mathcal{C})$ where $\mathcal{C}$ is the covariance operator. Then we are interested in a target Π such that:

$$\frac{d\Pi}{d\Pi_0} \propto \exp\left(-\Phi(q)\right), \tag{2.56}$$

where $\Phi(q) : \mathcal{H} \mapsto \mathbb{R}$. It turns out that the target distribution can be expressed as:

$$\Pi(x) \propto \exp\left(-\Phi(q) - \frac{1}{2}\left\langle q, \mathcal{C}^{-1} q \right\rangle\right), \tag{2.57}$$

for $q \in \mathcal{H}$. We set $L := \mathcal{C}^{-1}$. Then, for $x \in \mathcal{H}$, consider the following Hamiltonian which can be

understood as the corresponding version of (2.37) on $\mathcal{H} \times \mathcal{H}$:

$$\mathsf{H}(q,p) = \Phi(q) + \frac{1}{2}\langle q, Lx\rangle + \frac{1}{2}\langle p, \mathcal{M}p\rangle\,, \tag{2.58}$$

where $\mathcal{M}$ is a a user-specified *mass operator* which has to be well-defined covariance operator on $\mathcal{H}$. That is, $p \sim \mathcal{N}(0, \mathcal{M})$. We set $\mathcal{M} = L$ so that $\mathcal{C}^{-1} = \mathcal{M} = L$. Upon this specification, same as before, the Hamiltonian dynamics is given by:

$$\begin{cases} \frac{dq}{dt} = \nabla_p \mathsf{H}(q,p) = p, \\ \frac{dp}{dt} = -\nabla_q \mathsf{H}(q,p) = -q - \mathcal{C}\nabla\Phi(q). \end{cases} \tag{2.59}$$

here we omit dependency on t from notations for the sake of simplicity. Again, we can split (2.59) into the followings:

$$\mathsf{H}^{(1)}(q) := \Phi(q),\ \ \mathsf{H}^{(2)}(q,p) := \frac{1}{2}\langle q, Lq\rangle + \frac{1}{2}\langle p, Lp\rangle\,. \tag{2.60}$$

As for $\mathsf{H}^{(1)}(q)$, the corresponding dynamics are $\frac{dq}{dt} = \nabla_p \mathsf{H}^{(1)}(q) = 0$ and $\frac{dp}{dt} = -\nabla_q \mathsf{H}^{(1)}(q) = -\mathcal{C}\nabla\Phi(q)$. Also, as for $\mathsf{H}^{(2)}(q,p)$, we have that $\frac{dq}{dt} = \nabla_v \mathsf{H}^{(2)}(q,p) = p$ and $\frac{dp}{dt} = -\nabla_x \mathsf{H}^{(2)}(q,p) = -q$. Notice that both equations can be solved analytically. Then the corresponding solution operators are given by:

$$\Theta_t(q,p) = (q, p - t\mathcal{C}\nabla\Phi(q))\,, \tag{2.61}$$

$$\tilde{\Theta}_t(q,p) = (\cos(t)q + \sin(t)p, -\sin(t)q + \cos(t)p)\,, \tag{2.62}$$

and thus the numerical integrator is given by:

$$\Psi_\epsilon^{\circ(k+1)} = \Psi_\epsilon^{\circ(k)} \circ \Psi_\epsilon^{(1)}, \tag{2.63}$$

where we have defined $(q_{k+1}, p_{k+1}) = \Psi_\epsilon^{(1)} := \Theta_{\epsilon/2} \circ \tilde{\Theta}_\epsilon \circ \Theta_{\epsilon/2}(q_k, p_k)$. Notice that, with the choice $\cos(\epsilon^\star) = \frac{1-\epsilon^2/4}{1+\epsilon^2/4}$ for $\epsilon > 0$ (see Beskos et al. (2013a)), the corresponding numerical integrator can be alternatively expressed as:

$$\begin{cases} p_{k/2} = p_0 - \frac{\epsilon}{2}\frac{q_0+q_k}{2} - \frac{\epsilon}{2}\mathcal{C}\nabla\Phi(q_0), \\ q_k = q_0 + \epsilon p_{k/2}, \\ p_k = p_{k/2} - \frac{\epsilon}{2}\frac{q_0+q_k}{2} - \frac{\epsilon}{2}\mathcal{C}\nabla\Phi(q_k), \end{cases} \tag{2.64}$$

which can be understood as a semi-implicit-type integrator of the leap-frog. Also, in the same manner, we define the corresponding acceptance probability:

$$\alpha^{\mathsf{H}}(q_L, -p_L) := \min\left(1, \frac{\exp(-\mathsf{H}(q_L, -p_L))}{\exp(-\mathsf{H}(q_0, p_0))}\right), \tag{2.65}$$

and we summarise the algorithm as follows.

Algorithm 6 HMC on Hilbert spaces (Beskos et al., 2011, 2013a).

i) Draw p_n from $\mathcal{N}(0, \mathcal{M})$ given q_n.

ii) Obtain $(q^*_{n+1}, p^*_{n+1}) = \Psi^{\circ(L)}_\epsilon(q_n, p_n)$ by iterating (2.63) $L-1$ times.

iii) Set $q_{n+1} = q^*_{n+1}$ w.p.

$$\alpha^{\mathsf{H}}(q^*_{n+1}, -p^*_{n+1}) = \min\left(1, \frac{\exp\left(-\mathsf{H}(q^*_{n+1}, -p^*_{n+1})\right)}{\exp\left(-\mathsf{H}(q_n, p_n)\right)}\right),$$

otherwise set $q_{n+1} = q_n$.

iv) Discard p^*_{n+1}.

v) Repeat from the first step to the fourth step sufficiently large n times.

To see validity of Algorithm 6, consider the Gaussian product measure on $\mathcal{H} \times \mathcal{H}$ via the change of measure such that $Q_0 := \mathcal{N}(0, \mathcal{C}) \times \mathcal{N}(0, \mathcal{C})$ so that:

$$Q(dx, dv) := \exp\left(-\Phi(q)\right) Q_0(dq, dv).$$

Also we define the sequence of probability measures as follows:

$$Q^i := Q^{i-1} \circ \Psi^{-i},$$

for $1 \leq i \leq L$ and set:

$$g(q) := -\mathcal{C}^{1/2} \nabla \Phi(q), \tag{2.66}$$

for $q \in \mathcal{H}$. Notice that we need the assumption which ensures that $\mathcal{C}\nabla\Phi(q)$ *is an element of the Cameron-Martin space of the Gaussian measure* Π_0 *for any* $q \in \mathcal{H}$ *w.p.1. under the measure* Π_0. Besides, we need the following lemma.

Lemma 7. *Let* $\Pi_0 = \mathcal{N}(0, \mathcal{C})$ *on* $\mathcal{H}$ *and set* $T(q) = q + \mathcal{C}^{1/2} q^0$, $q \in \mathcal{H}$ *and where* $q^0 \in \mathcal{H}$ *is a constant. Then we have that*

$$\frac{d\left\{\Pi_0 \circ T^{-1}\right\}}{d\left\{\Pi_0\right\}}(q) = \exp\left(-\frac{1}{2} \mid q^0 \mid^2 + \left\langle q_0, C^{-1/2} q \right\rangle\right).$$

Proof. Proposition 1.17 of Da Prato (2006) ensures that the measure $\Pi_0 \circ T^{-1}$ is $\mathcal{N}(\mathcal{C}^{1/2} q^0, \mathcal{C})$. Since $\mathcal{C}^{1/2} q^0$ is an element of the Cameron-Martin space of $\mathcal{C}$, $\Pi_0 \circ T^{-1}$ and Π_0 are equivalent so that $\frac{d\left\{\Pi_0 \circ T^{-1}\right\}}{d\left\{\Pi_0\right\}}(q) = \exp\left(-\frac{1}{2} \mid q^0 \mid^2 + \left\langle q_0, C^{-1/2} q \right\rangle\right)$ follows from Theorem 2.8 of Da Prato (2006). □

For the sake of completeness, we provide a self-contained proof, based on Beskos et al. (2013a).

Lemma 8. *We have that:*

$$\frac{dQ^i}{dQ}(q_i,p_i) = \frac{dQ^{i-1}}{dQ_0}(q_{i-1},p_{i-1}) \times G(q_i,p_i) \times G(q_{i-1},p_{i-1} + \frac{\epsilon}{2}\mathcal{C}^{1/2}g(q_{i-1})),$$

where we have defined:

$$G(q,p) := \exp\left(\left\langle \frac{\epsilon}{2}g(q), \mathcal{C}^{-1/2}p\right\rangle - \frac{1}{2}\left|\frac{\epsilon}{2}g(q)\right|^2\right) = \frac{d\left\{Q_0 \circ \Theta^{-1}_{\epsilon/2}\right\}}{dQ_0}(q,p), \tag{2.67}$$

where $g(x)$ is defined in (2.66).

Proof. From the definition, first we have $Q^i = Q^{i-1} \circ \Theta^{-1}_{\epsilon/2} \circ \tilde{\Theta}^{-1}_{\epsilon/2} \circ \Theta^{-1}_{\epsilon/2}$. Then we get:

$$\begin{aligned}
\frac{dQ^i}{dQ_0}(q_i,p_i) &= \frac{d\left\{Q^{i-1} \circ \Theta^{-1}_{\epsilon/2} \circ \tilde{\Theta}^{-1}_{\epsilon^\star} \circ \Theta^{-1}_{\epsilon/2}\right\}}{dQ_0}(q_i,p_i), \\
&= \frac{d\left\{Q^{i-1} \circ \Theta^{-1}_{\epsilon/2} \circ \tilde{\Theta}^{-1}_{\epsilon^\star} \circ \Theta^{-1}_{\epsilon/2}\right\}}{d\left\{Q_0 \circ \Theta^{-1}_{\epsilon/2}\right\}}(q_i,p_i) \times \frac{d\left\{Q_0 \circ \Theta^{-1}_{\epsilon/2}\right\}}{dQ_0}(q_i,p_i), \\
&= \frac{d\left\{Q^{i-1} \circ \Theta^{-1}_{\epsilon/2} \circ \tilde{\Theta}^{-1}_{\epsilon^\star}\right\}}{dQ_0}\left(\Theta^{-1}_{\epsilon/2}(q_i,p_i)\right) \times G(q_i,p_i),
\end{aligned}$$

here we used Lemma Lemma 7 with $q_0 = g(q)$. Notice that $\Theta^{-1}_{\epsilon/2} \circ \tilde{\Theta}^{-1}_{\epsilon^\star}(q_i,p_i) = \Theta_{\epsilon/2}(q_{i-1},p_{i-1})$ holds by the construction. As a result, we obtain:

$$\begin{aligned}
\frac{d\left\{Q^{i-1} \circ \Theta^{-1}_{\epsilon/2} \circ \tilde{\Theta}^{-1}_{\epsilon^\star}\right\}}{dQ_0}\left(\Theta^{-1}_{\epsilon/2}(q_i,p_i)\right) &= \frac{d\left\{Q^{i-1} \circ \Theta^{-1}_{\epsilon/2} \circ \tilde{\Theta}^{-1}_{\epsilon^\star}\right\}}{\left\{dQ_0 \circ \tilde{\Theta}^{-1}_{\epsilon^\star}\right\}}\left(\Theta^{-1}_{\epsilon/2}(q_i,p_i)\right), \\
&= \frac{d\left\{Q^{i-1} \circ \Theta^{-1}_{\epsilon/2}\right\}}{dQ_0}\left(\Theta_{\epsilon/2}(q_{i-1},p_{i-1})\right), \\
&= \frac{d\left\{Q^{i-1} \circ \Theta^{-1}_{\epsilon/2}\right\}}{d\left\{Q_0 \circ \Theta^{-1}_{\epsilon/2}\right\}}\left(\Theta_{\epsilon/2}(q_{i-1},p_{i-1})\right) \times \frac{d\left\{Q_0 \circ \Theta^{-1}_{\epsilon/2}\right\}}{dQ_0}\left(\Theta_{\epsilon/2}(q_{i-1},p_{i-1})\right), \\
&= \frac{dQ^{i-1}}{dQ_0}(q_{i-1},p_{i-1}) \times G\left(\Theta_{\epsilon/2}(q_{i-1},p_{i-1})\right), \\
&= \frac{dQ^{i-1}}{dQ_0}(q_{i-1},p_{i-1}) \times G\left(q_{i-1},p_{i-1} + \frac{\epsilon}{2}\mathcal{C}^{1/2}g(q_{i-1})\right).
\end{aligned}$$

here notice that $Q_0 \circ \tilde{\Theta}^{-1}_{\epsilon^\star} = Q_0$, the expression in (2.67) and the construction of $\Theta_{\epsilon/2}$. Therefore, we have that:

$$\frac{dQ^i}{dQ_0}(q_i,p_i) = \frac{dQ^{i-1}}{dQ_0}(q_{i-1},p_{i-1}) \times G(q_i,p_i) \times G\left(q_{i-1},p_{i-1} + \frac{\epsilon}{2}\mathcal{C}^{1/2}g(q_{i-1})\right).$$

□

Critically, lemma Lemma 8 gives rise to:

$$\frac{dQ^L}{dQ_0}(q_L, p_L) = \frac{dQ}{dQ_0}(q_0, p_0) \prod_{i=1}^{L} G(q_i, p_i) G\left(q_{i-1}, p_{i-1} + \frac{\epsilon}{2}\mathcal{C}^{1/2} g(q_{i-1})\right).$$

Moreover, define:

$$\begin{aligned} p_{i-1}^{-} &:= \mathfrak{proj}^{\mathfrak{p}} \circ \Theta_{\epsilon/2}(q_{i-1}, p_{i-1}) = p_{i-1} + \frac{\epsilon}{2}\mathcal{C}^{1/2} g(q_{i-1}), \\ p_{i-1}^{+} &:= \mathfrak{proj}^{\mathfrak{p}} \circ \tilde{\Theta}_{\epsilon^\star} \circ \Theta_{\epsilon/2}(q_{i-1}, p_{i-1}) = p_i - \frac{\epsilon}{2}\mathcal{C}^{1/2} g(q_i). \end{aligned}$$

Using these, we can calculate as follows (recall $\mathcal{C}^{-1} = L = \mathcal{M}$):

$$\begin{aligned} &\log\left\{G(q_i, p_i) G\left(q_{i-1}, p_{i-1} + \frac{\epsilon}{2}\mathcal{C}^{1/2} g(q_{i-1})\right)\right\} = \\ &= \left\langle \frac{\epsilon}{2} g(q_i), \mathcal{C}^{-1/2} p_i \right\rangle - \frac{1}{2}\left|\frac{\epsilon}{2} g(q_i)\right|^2 + \left\langle \frac{\epsilon}{2} g(q_{i-1}), \mathcal{C}^{-1/2} p_{i-1} \right\rangle - \frac{1}{2}\left|\frac{\epsilon}{2} g(q_{i-1})\right|^2, \\ &= \frac{1}{2}\left\langle p_i, L p_i \right\rangle + \frac{1}{2}\left\langle p_i^+, L p_i^+ \right\rangle + \frac{1}{2}\left\langle p_{i-1}, L p_{i-1} \right\rangle - \frac{1}{2}\left\langle p_{i-1}^+, L p_{i-1}^+ \right\rangle, \\ &= \frac{1}{2}\left\langle q_i, L q_i \right\rangle + \frac{1}{2}\left\langle p_i, L p_i \right\rangle - \frac{1}{2}\left\langle q_{i-1}, L q_{i-1} \right\rangle - \frac{1}{2}\left\langle p_{i-1}, L p_{i-1} \right\rangle, \end{aligned}$$

so that $\log \sum_{i=1}^{L} G(q_i, p_i) G\left(q_{i-1}, p_{i-1} + \frac{\epsilon}{2}\mathcal{C}^{1/2} g(q_{i-1})\right) = \mathsf{H}(q_L, p_L) - \Phi(q_L) - \frac{1}{2}\langle q_0, L p_0 \rangle - \frac{1}{2}\langle p_0, L p_0 \rangle$. Since $\frac{dQ}{dQ_0}(q_0, p_0) = \exp\left(-\Phi(q_0)\right) = \exp\left(\frac{1}{2}\langle q_0, L q_0 \rangle + \frac{1}{2}\langle p_0, \mathcal{M} p_0 \rangle - \mathsf{H}(q_0, p_0)\right)$, we finally obtain the following proposition.

Proposition 21. *Q^L is absolutely continuous w.r.t. Q_0, and:*

$$\frac{dQ^L}{dQ_0}(q_L, p_L) = \exp\left(\mathsf{H}(q_L, p_L) - \mathsf{H}(q_0, p_0) - \Phi(q_L)\right).$$

Roughly speaking, Proposition 21 implies that Metropolis acceptance ratio can be well defined on the concerning Hilbert space. Next, we check the Markov kernel induced by the advanced HMC is indeed reversible. To do so, we heuristically assume that $\{\Psi_\epsilon^{\circ(k)}\}$ for $k \in \{0, \cdots, L\}$ preserves volume on $\mathcal{H}$. This is true, for instance, $\mathcal{H} = \mathbb{R}^d$.

Proposition 22. *The Markov kernel induced by Algorithm 6 leaves the target invariant.*

Proof. Assume that $(q_0, p_0) \sim Q$. Then for the next position q' and $\mathfrak{proj}^{\mathfrak{q}} \circ \Psi_\epsilon^{\circ(L)}(q_0, p_0) = q_L$, we have that:

$$q' = \mathbb{I}\left\{u \leq \alpha^{\mathsf{H}}(q_0, p_0)\right\} q_L + \mathbb{I}\left\{u > \alpha^{\mathsf{H}}(q_0, p_0)\right\} q_0,$$

where $u \sim Unif(0, 1)$ and notice that $(q_0, p_0) = \Psi_\epsilon^{-\circ(L)}(q_L, p_L)$. Let $f \in \mathcal{B}_b(\mathcal{H})$. Since we have assumed that $(q_0, p_0) \sim Q$, we want to show that $\mathbb{E}\left[f(q_0)\right] = \mathbb{E}\left[f(q')\right]$. Taking the expectation w.r.t. u gives rise to:

$$\mathbb{E}\left[f(q')\right] = \mathbb{E}\left[f(q_L)\alpha^{\mathsf{H}}(q_0, p_0)\right] - \mathbb{E}\left[f(q_0)\alpha^{\mathsf{H}}(q_0, p_0)\right] + \mathbb{E}\left[f(q_0)\right]. \tag{2.68}$$

From Proposition 21, we can show that:

$$\begin{aligned}
\mathbb{E}\left[f(q_L)\alpha^{\mathsf{H}}(q_0,p_0)\right] &= \mathbb{E}_{Q^L}\left[f(q_L)\alpha^{\mathsf{H}}\left(\Psi_\epsilon^{-\circ(L)}(q_L,p_L)\right)\right], \\
&= \mathbb{E}_{Q_0}\left[f(q_L)\alpha^{\mathsf{H}}\left(\Psi_\epsilon^{-\circ(L)}(q_L,p_L)\right)\exp\left(\mathsf{H}(q_L,p_L)-\mathsf{H}(q_0,p_0)-\Phi(q_L)\right)\right], \\
&= \mathbb{E}_{Q_0}\left[f(q_L)\max\left\{1,\exp\left(\Delta\mathsf{H}\left(\Psi_\epsilon^{-\circ(L)}(q_L,p_L)\right)\right)\right\}\exp\left(-\Phi(q_L)\right)\right], \\
&= \mathbb{E}_{Q}\left[f(q_L)\max\left\{1,\exp\left(\Delta\mathsf{H}\left(\Psi_\epsilon^{-\circ(L)}(q_L,p_L)\right)\right)\right\}\right], \\
&= \mathbb{E}_{Q}\left[f(q_L)\max\left\{1,\exp\left(\Delta\mathsf{H}\left(\Psi_\epsilon^{-\circ(L)}(q_L,-p_L)\right)\right)\right\}\right].
\end{aligned}$$

Recall that for S in (2.43), we have $S\circ\Psi_\epsilon^{\circ(L)} = \left(\Psi_\epsilon^{-\circ(L)}\right)\circ S$, and this yields:

$$\begin{aligned}
\Delta\mathsf{H}\left(\Psi_\epsilon^{-\circ(L)}(q_L,-p_L)\right) &= \Delta\mathsf{H}\left(S\circ\Psi_\epsilon^{\circ(L)}(q_L,p_L)\right), \\
&= \mathsf{H}\left(S(q_L,p_L)\right) - \mathsf{H}\left(S\circ\Psi_\epsilon^{\circ(L)}(q_L,p_L)\right), \\
&= \mathsf{H}(q_L,p_L) - \mathsf{H}\left(\Psi_\epsilon^{\circ(L)}(q_L,p_L)\right) = -\Delta\mathsf{H}(q_L,p_L).
\end{aligned}$$

where we have defined $\Delta\mathsf{H}(q,p) := \mathsf{H}\left(\Psi_\epsilon^{\circ(L)}(q_L,p_L)\right) - \mathsf{H}(q,p)$, and used $S = \Psi_\epsilon^{\circ(L)}\circ S\circ\Psi_\epsilon^{\circ(L)}$ and $\mathsf{H}\circ S = \mathsf{H}$ since the kinetic function is $\frac{1}{2}\langle p, Lp\rangle$ i.e., quadratic in p. Hence,

$$\begin{aligned}
\mathbb{E}\left[f(q_L)\alpha^{\mathsf{H}}(q_0,p_0)\right] &= \mathbb{E}_{Q}\left[f(q_L)\max\left\{1,\exp\left(\Delta\mathsf{H}\left(\Psi_\epsilon^{-\circ(L)}(q_L,-p_L)\right)\right)\right\}\right], \\
&= \mathbb{E}_{Q}\left[f(q_L)\alpha^{\mathsf{H}}(q_L,p_L)\right] = \mathbb{E}\left[f(q_0)\alpha^{\mathsf{H}}(q_0,p_0)\right],
\end{aligned}$$

here again notice that we have assumed $(q_0,p_0)\sim Q$, so that the claim follows from (2.68). □

Notice that the practical application of Algorithm 6 requires of course to replace H, Π_0 and Φ by finite-dimensional approximations. That is, Algorithm 6 is the d-dimensional proxy of the $\mathcal{H}$ valued HMC on $\mathbb{R}^d$ in practice. Beskos et al. (2011) show that the d-dimensional proxy of Algorithm 6 converges to the algorithm on $\mathcal{H}$ as $d\uparrow\infty$. Critically, Beskos et al. (2011, Theorem 4.1) also ensure that, in contrast with the standard HMC (Algorithm 5), the d-dimensional proxy of $\alpha^{\mathsf{H}}(\cdot,\cdot)$ does not degenerate into 0 as d increases, for fixed time-step ϵ in the integrator. This property is often called *mesh-free* in the sense that, when applying the algorithms on a computer, mixing times do not deteriorate even though the number of mesh points of the approximation of the infinite-dimensional increase, see Cotter et al. (2013) for instance.

3 Sequential Monte Carlo

3.1 Introduction

Let $(E, \mathcal{E})$ be a measurable space. Assume that we are interested in a sequence of targets $\{\pi_n(x_{0:n})\}$ of increasing (w.r.t. n) dimension with each $\pi_n(x_{0:n})$ is defined on the product space $(E^n, \mathcal{E}^n) := (\prod_{p=0}^n E_p, \mathcal{E}^{\otimes n})$. Also, assume that the sequence of target probability densities $\{\pi_n(x_{0:n})\}$ is known up to the normalising constant, that is, we have that:

$$\pi_n(x_{0:n}) = \frac{\gamma_n(x_{0:n})}{Z_n}, \tag{3.1}$$

$$Z_n := \int \gamma_n(x_{0:n}) dx_{0:n}, \tag{3.2}$$

and only $\gamma_n(x_{0:n})$ is known point-wise. Sequential Monte Carlo (SMC) is a general class of Monte Carlo methods which sample sequentially from such a sequence of targets $\{\pi_n(x_{0:n})\}$. We refer to Doucet and Johansen (2009); Naesseth et al. (2019); Douc et al. (2014) as a general reference of SMC. The aim of this section is to provide some basic and detailed results of SMC methods which will appear implicitly and explicitly in the rest of the thesis. For sake of simplicity, we will assume that, depending on the situation, (3.1) admits the densities w.r.t. the appropriate reference measure.

3.2 Basics of Sequential Monte Carlo

Roughly speaking, SMC consists of sequential importance sampling and resampling. Instead of sampling from $\gamma_n(x_{0:n})$ directly, assume that one can obtain samples from the *importance density* $q(x_{0:n})$ which has the following structure:

$$\begin{aligned} q(x_{0:n}) &= q(x_{0:n-1}) q(x_n \mid x_{0:n-1}), \\ &= q(x_0) \prod_{i=1}^n q(x_i \mid x_{1:i-1}), \end{aligned} \tag{3.3}$$

see Appendix F for a brief explanation of importance sampling. Note that this structure of the importance density implies that one can sample $x_1^{(i)} \sim q(x_1)$ at time 1, $x_2^{(2)} \sim q(x_2 \mid x_1^{(i)})$ at time 2 and then $x_n^{(i)} \sim q(x_n \mid x_{0:n-1}^{(i)})$ at time $n \geq 3$. Also, we require that whenever $\gamma_n(x_{0:n}) > 0$, $q(x_{0:n}) > 0$ as well. That is, roughly speaking, the law induced by $\gamma_n(x_{0:n})$ has to be absolutely continuous w.r.t. the one induced by $q(x_{0:n})$. Then, as the Radon-Nikodym derivative, one can define the *unnormalised weights* as follows:

$$\begin{aligned} w_n(x_{0:n}) &:= \frac{\gamma_n(x_{0:n})}{q(x_{0:n})}, \\ &= \frac{\gamma_{n-1}(x_{0:n-1})}{q(x_{0:n-1})} \times \frac{\gamma_n(x_{0:n})}{q(x_n \mid x_{0:n-1}) \gamma_{n-1}(x_{0:n-1})}, \\ &= w_{n-1}(x_{0:n-1}) \times \alpha_n(x_{0:n}), \end{aligned} \tag{3.4}$$

where we have defined the *incremental importance weight* function $\alpha_n(x_{0:n})$:

$$\alpha_n(x_{0:n}) := \frac{\gamma_n(x_{0:n})}{q(x_n \mid x_{0:n-1})\gamma_{n-1}(x_{0:n-1})}. \tag{3.5}$$

Then iterating this procedure gives rise to the recursive equation for $w_n(x_{0:n})$ such that:

$$w_n(x_{0:n}) = \prod_{k=0}^{n} \alpha_k(x_{0:k}). \tag{3.6}$$

Also, it is clear to see that, for any $f \in \mathcal{B}_b(E^n)$, we have that:

$$\mathbb{E}_\pi\left[f(x_{0:n})\right] = \frac{\mathbb{E}_q\left[f(x_{0:n})w_n(x_{0:n})\right]}{\mathbb{E}_q\left[w_n(x_{0:n})\right]},$$

holds for any n. Then *Sequential Importance Sampling* (SIS) (Kong et al., 1994) at time n might be done as follows. Assume that one has a set of approximations of $\gamma_{n-1}(x_{0:n-1})$, say $(x_{0:n-1}^{(i)}, W_{n-1}^{(i)})_{i=1}^N$. Then, SIS can be implemented: (1) *propagate* $x_n^{(i)}$ from the importance density $q(\cdot \mid x_{0:n-1}^{(i)})$ and set $x_{0:n}^{(i)} := (x_n^{(i)}, x_{0:n-1}^{(i)})_{i=1}^N$. (2) Next, *correct* the unnormalised weights via the $\alpha_n(x_{0:n})$ in (3.5) then obtain the *normalised weights* such as:

$$W_n^{(i)} := \frac{w_n(x_{0:n}^{(i)})}{\sum_{j=1}^N w_n(x_{0:n}^{(j)})}. \tag{3.7}$$

The SIS algorithm can be summarised as follows.

Algorithm 7 Sequential Importance Sampling (SIS) (Kong et al., 1994)

Assume that at time $n-1$, one has $(x_{0:n-1}^{(i)}, W_{n-1}^{(i)})_{i=1}^N$ targeting $\pi_{n-1}(dx_{0:n-1})$ and time n,

i) Propagate particles $\{x_n^{(i)}\}_{i=1}^N$ via sampling from $q(\cdot \mid x_{0:n-1}^{(i)})$.

ii) Correct unnormalised weights via $w_n(x_{0:n}^{(i)}) = w_{n-1}(x_{0:n-1}^{(i)})\alpha(x_{0:n}^{(i)})$ for $i = 1, \dots N$.

iii) Obtain normalised weights via $W_n^{(i)} = \frac{w_n(x_{0:n}^{(i)})}{\sum_{j=1}^N w_n(x_{0:n}^{(j)})}$ for $i = 1, \dots N$.

iv) Return to the first step.

In the literature, $x_{0:n}^{(i)}$ simulated from the importance density are often called *particles* and the collection of $(x_{0:n}^{(i)}, W_n^{(i)})_{i=1}^N$ is often called a *weighted particle system*, and this gives rise to the (weighted) empirical measure on E^n as follows:

$$\hat{\pi}_n(dx_{0:n}) := \sum_{i=1}^N W_n^{(i)} \delta_{x_{0:n}^{(i)}}(dx_{0:n}), \tag{3.8}$$

where $\delta_{x_{0:n}^{(i)}}$ denotes the Dirac mass located at $x_{0:n}^{(i)}$. Also, notice that:

$$\begin{aligned}\int \alpha_n(x_{0:n})\pi_{n-1}(x_{0:n-1})q(x_n \mid x_{0:n-1})dx_{0:n} &= \int \frac{\gamma_n(x_{0:n})\pi_{n-1}(x_{0:n-1})q(x_n \mid x_{0:n-1})}{\gamma_{n-1}(x_{0:n-1})q(x_n \mid x_{0:n-1})}dx_{0:n}, \\ &= \int \frac{\gamma_n(x_{0:n})\frac{\gamma_{n-1}(x_{0:n-1})}{Z_{n-1}}}{\gamma_{n-1}(x_{0:n-1})}dx_{0:n} = \frac{Z_n}{Z_{n-1}},\end{aligned}$$

and thus, this gives rise to the estimate of $\frac{Z_n}{Z_{n-1}}$ such that:

$$\widehat{\frac{Z_n}{Z_{n-1}}} := \sum_{i=1}^{N} W_{n-1}^{(i)}\alpha_n(x_{0:n}^{(i)}). \tag{3.9}$$

Critically, as the by-product of (3.9), we obtain the estimate of Z_n:

$$\hat{Z}_n := \hat{Z}_0 \prod_{p=1}^{n}\left(\sum_{i=1}^{N} W_{p-1}^{(i)}\alpha_p(x_{0:p}^{(i)})\right). \tag{3.10}$$

Apparently, one can do this procedure recursively and thus the cost of SIS is constant, that is $\mathcal{O}(N)$, w.r.t. the time index n. Therefore, SIS itself is an on-line method. Although SIS presented above is sometimes useful, it has the serious problem that is known as *degeneracy of the weights*. Roughly speaking, the conditional variance of $w(x_{0;n})$ will increase as the time index n increases. Indeed, we can show that, by using the law of total variance,

$$\begin{aligned}\mathbb{V}\left[w_n(x_{0:n}) \mid \mathcal{F}_{n-1}\right] &= \mathbb{V}\left[w_{n-1}(x_{0:n-1})\frac{\gamma_n(x_{0:n})}{q_n(x_n \mid x_{0:n-1})\gamma_{n-1}(x_{0:n-1})} \mid \mathcal{F}_{n-1}\right], \\ &= \mathbb{E}\left[\mathbb{V}\left[w_{n-1}(x_{0:n-1})\frac{\gamma_n(x_{0:n})}{q_n(x_n \mid x_{0:n-1})\gamma_{n-1}(x_{0:n-1})} \mid \mathcal{F}_{n-1}\right]\right] \\ &+ \mathbb{V}\left[\mathbb{E}\left[w_{n-1}(x_{0:n-1})\frac{\gamma_n(x_{0:n})}{q_n(x_n \mid x_{0:n-1})\gamma_{n-1}(x_{0:n-1})} \mid \mathcal{F}_{n-1}\right]\right], \\ &\geq \mathbb{V}\left[w_{n-1}(x_{0:n-1})\mathbb{E}\left[\frac{\gamma_n(x_{0:n})}{q_n(x_n \mid x_{0:n-1})\gamma_{n-1}(x_{0:n-1})} \mid \mathcal{F}_{n-1}\right]\right], \\ &= \mathbb{V}\left[w_{n-1}(x_{0:n-1})\right],\end{aligned}$$

holds for any $n \geq 1$, where $\mathcal{F}_n$ is the natural filtration generated by a particle system at time n. This problem would have the effect that all but one of the weights decreases to zero at the end of the day, and all emphasis is thus put on one of the particles. Also it can be shown that the choice $q(x_n \mid x_{0:n-1}) = \pi_n(x_n \mid x_{0:n-1})$ is the best proposal in the sense that this choice minimises the conditional variance of $w(x_{0:n})$, that is $\mathbb{V}[w_n(x_{0:n}) \mid \mathcal{F}_{n-1}] = 0$.

Then *resampling* is a common way to overcome this problem. The idea behind resampling is very simple and intuitive. Recall that, at the moment, we do not do sampling from $\pi_n(dx_{0:n})$ but from $q_n(dx_{0:n})$. Therefore, in order to obtain particles form the target, one can make use of the empirical measure $\hat{\pi}_n(dx_{0:n})$ constructed by a particle system $(x_{0:n}^{(i)}, W_n^{(i)})_{i=1}^N$.That is, one can just select $x_{0:n}^{(i)}$ w.p. $W_n^{(i)}$ and this procedure is equivalent to associate the number of offspring $N_n^{(i)}$ with the $i-$th

particles $x_{0:n}^{(i)}$ based on their weights $W_n^{(i)}$. Namely, one just needs to draw N samples $N_n^{(i)}$ from the multinomial distribution with parameter vector $(N, W_n^{(i)})_{i=1}^N$ and this gives rise to a new particle system $(\tilde{x}_{0:n}^{(i)}, \frac{1}{N})_{i=1}^N$. This resampling scheme is commonly reoffered as *Multinomial Resampling*.

However, mathematically speaking, the only requirement for a resampling system is that it has to be unbiased in the following sense. Let $N_n^{(i)}$ denote the number of offspring which is associated to the particle $x_{0:n}^{(i)}$. Then, for any n, $\mathbb{E}\left[N_n^{(i)} \mid \mathcal{F}_n\right] = N$ has to hold.Then, after resampling, a equally weighted particle system $(\tilde{x}_{0:n}^{(i)}, \frac{1}{N})_{i=1}^N$ constructs the unweighted empirical measure:

$$\tilde{\pi}_n(dx_{0:n}) := \sum_{i=1}^N \frac{N_n^{(i)}}{N} \delta_{\tilde{x}_{0:n}^{(i)}}(dx_{0:n}), \tag{3.11}$$

and this is indeed unbiased.

In practice, *Systematic Resampling* (Kitagawa, 1996) might be the most popular resampling method due to its stability and computational efficiency. See Douc and Cappé (2005) for the theoretical comparison of various resampling approaches. Critically, resampling will remove less important particles $\{x_{0:n}^{(i)}\}_{i=1}^N$ measured by weights $\{W_n^{(i)}\}_{i=1}^N$ and rejuvenate ones according to their significance.

Algorithm 8 Systematic Resampling (Kitagawa, 1996)

i) Set $C = 0$ and $j = 1$. Draw $u \sim \mathcal{U}[0, 1]$.

ii) For $i = 1 \cdots N$, set $C = W_n^{(i)}$. Then while $\frac{u+j-1}{N} \leq C$, set $\tilde{x}_{0:n}^{(j)} = x_{0:n}^{(i)}$ and $j = i + 1$.

Now we are ready to describe a generic SMC algorithm. SMC consists of SIS and resampling, and which can be decomposed into the following 3 steps, *mutation*, *correction* and *selection*. The mutation step and the correction steps are same as SIS. After the correction step, one has a weighted particle system $\left(x_{0:n}^{(i)}, W_n^{(i)}\right)_{i=1}^N$. Then, at selection step, one obtains an equally particle system $\left(\tilde{x}_{0:n}^{(i)}, \frac{1}{N}\right)_{=1}^N$ via some resampling methods. Note that now $\left\{x_{0:n}^{(i)}\right\}_{i=1}^N := \left\{\tilde{x}_{0:n-1}^{(i)}, x_n^{(i)}\right\}_{i=1}^N$ is approximately distributed according to $\pi_{n-1}(x_{0:n-1})q_n(x_n \mid x_{n-1})$ so that the corresponding importance weights in this case are simply equal to $\alpha(x_{0:n}^{(i)})$.

Algorithm 9 Sequential Monte Carlo

Assume that at time $n-1$, one has an *equally particle system* $\left(\tilde{x}_{0:n-1}^{(i)}, \frac{1}{N}\right)_{i=1}^N$ of the target $\pi(dx_{0:n-1})$.

i) Propagate particles $\{x_n^{(i)}\}_{i=1}^N$ via sampling from $q(\cdot \mid \tilde{x}_{0:n-1}^{(i)})$ and set $x_{0:n}^{(i)} \leftarrow \{x_n^{(i)}, \tilde{x}_{0:n-1}^{(i)}\}_{i=1}^N$.

ii) Correct unnormalised weights via $w_n(x_{0:n}^{(i)}) = \alpha(x_{0:n}^{(i)})$ for $i = 1, \ldots N$.

iii) Obtain normalised weights via $W_n^{(i)} = \frac{w_n^{(i)}(x_{0:n})}{\sum_{j=1}^N w_n^{(j)}(x_{0:n})}$ for $i = 1, \ldots N$.

iv) Do resampling $\left\{x_n^{(i)}\right\}_{i=1}^N$ w.p. $W_n^{(i)}$ to obtain equally weighted particle system $(\tilde{x}_{0:n}^{(i)}, \frac{1}{N})_{i=1}^N$.

v) Return to the first step.

At any time $n \geq 0$, Algorithm 9 constructs (3.8) and:

$$\tilde{\pi}_n(dx_{0:n}) := \frac{1}{N}\sum_{i=1}^{N} \delta_{\tilde{x}_{0:n}^{(i)}}(dx_{0:n}).$$

Also, the estimate of Z_n in (3.10) becomes:

$$\begin{aligned}\hat{Z}_n &:= \hat{Z}_0 \prod_{p=1}^{n} \left(\sum_{i=1}^{N} W_{p-1}^{(i)} \alpha_p(\tilde{x}_{0:p}^{(i)}) \right), \\ &= \hat{Z}_0 \prod_{p=1}^{n} \left(\frac{1}{N} \sum_{i=1}^{N} \alpha_p(\tilde{x}_{0:p}^{(i)}) \right).\end{aligned}$$

Although resampling has theoretical justification and could improve performance of SMC, as pointed out theoretically (Del Moral et al., 2010) and experimentally (Kantas et al., 2015), such path-space method (targetting $\pi_n(dx_{0:n})$) may suffer from the particle *path degeneracy problem* which is also well-known in the SMC literature. Roughly speaking, as $n \to \infty$, resampled particles $\{\tilde{x}_{0:n}^{(i)}\}_{i=1}^N$ might share a common *ancestor* due to the successive resampling steps. Also, resampling introduces additional noise. Therefore, resampling should be understood as the price to pay to obtain reasonable approximations of $\pi(dx_{0:n})$ for the long run, at the cost of putting instability for the short run.

Therefore, it would be desirable that one does resampling only when the obtained particle system $(x_{0:n}^{(i)}, W_n^{(i)})_{i=1}^N$ is inefficient in some sense. Then common choice of the quantity to monitor effectiveness of SMC is the *Effective Sample Size* (ESS) (Liu and Chen, 1998; Kong, 1992; Kong et al., 1994), which is defined as follows:

$$ESS := \frac{1}{\sum_{i=1}^{N} (W_n^{(i)})^2}. \tag{3.12}$$

Since $\{W_n^{(i)}\}_{i=1}^N$ are normalised, the ESS is a positive continuous variable taking values in $(1, N)$. In the context of importance sampling, the ESS (at time n) can be understood as the number of independent samples generated directly from the target $\pi_n(dx_{0:n})$, which yields the same efficiency in the estimation obtained by an approximation from $q_n(dx_{0:n})$, and thus large values of the ESS can be understood that the obtained particle system $\{x_{0:n}^{(i)}, W_n^{(i)}\}_{i=1}^N$ well approximates the target $\pi_n(dx_{0:n})$ and vice versa. By using the ESS, the user can specify the threshold in advance, and when ESS is lower than it, one can do SMC dynamically. A typical threshold is $\frac{N}{2}$.

Algorithm 10 SMC with dynamic resampling

Assume that at time $n-1$, one has *weighted particle system* $(\tilde{x}_{0:n-1}^{(i)}, \tilde{W}_{n-1}^{(i)})_{i=1}^N$ of the target $\pi(dx_{0:n-1})$.

i) Propagate particles $\{x_n^{(i)}\}_{i=1}^N$ via sampling from $q_n(\cdot \mid \tilde{x}_{0:n-1}^{(i)})$ and set $x_{0:n}^{(i)} \leftarrow \{x_n^{(i)}, \tilde{x}_{0:n-1}^{(i)}\}_{i=1}^N$.

ii) Correct unnormalised weights via $w_n(x_{0:n}^{(i)}) = \tilde{W}_{n-1}^{(i)} \times \alpha_n(x_{0:n}^{(i)})$ for $i = 1, \dots N$.

iii) Obtain normalised weights via $W_n^{(i)} = \frac{w_n^{(i)}}{\sum_{j=1}^N w_n^{(j)}}$ for $i = 1, \dots N$.

iv) Calculate the ESS, $\frac{1}{\sum_{i=1}^N (W_n^{(i)})^2}$.

v) If the obtained ESS is smaller than a threshold, do resampling $\left\{x_n^{(i)}\right\}_{i=1}^N$ w.p. $W_n^{(i)}$ and set $(\tilde{x}_n^{(i)}, \tilde{W}_n^{(i)})_{i=1}^N \leftarrow (\tilde{x}_n^{(i)}, \frac{1}{N})_{i=1}^N$ to obtain equally weighted particle system $(\tilde{x}_{0:n}^{(i)}, \frac{1}{N})_{i=1}^N$.

vi) If it is not, set $(\tilde{x}_n^{(i)}, \tilde{W}_n^{(i)})_{i=1}^N \leftarrow (x_n^{(i)}, W_n^{(i)})_{i=1}^N$ to obtain weighted particle system $(\tilde{x}_{0:n}^{(i)}, \frac{1}{N})_{i=1}^N$.

vii) Return to the first step.

At any time $n \geq 1$, Algorithm 10 provides two empirical measures:

$$\begin{cases} \hat{\pi}_n(dx_{0:n}) & = \sum_{i=1}^N W_n^{(i)} \delta_{x_{0:n}^{(i)}}(dx_{0:n}), \\ \tilde{\pi}_n(dx_{0:n}) & = \sum_{i=1}^N \tilde{W}_n^{(i)} \delta_{\tilde{x}_{0:n}^{(i)}}(dx_{0:n}), \end{cases} \tag{3.13}$$

which are equal if no resampling occurred at time n, and:

$$\widehat{\frac{Z_n}{Z_{n-1}}} = \sum_{i=1}^N \tilde{W}_{n-1}^{(i)} \alpha_n(x_{0:n}^{(i)}). \tag{3.14}$$

Finally, we end this subsection by noting the useful technique called *logsumexp*. Recall that the normalised weight is given by $W_n^{(i)} := \frac{w_n^{(i)}}{\sum_{i=1}^N w_n^{(i)}}$. As $N \to \infty$, each $w_n^{(i)}$ will take really small values and this makes numerical instability in $W_n^{(i)}$ due to the term $1/\sum_{i=1}^N w_n^{(i)}$. In order to avoid such instability, one might prefer using $\log W_n^{(i)}$ to $W_n^{(i)}$, and obtain $W_n^{(i)} = \exp\left(\log W_n^{(i)}\right)$. That is, one might want to calculate:

$$\log W_n^{(i)} = \log\left(\frac{w_n^{(i)}}{\sum_{i=1}^N w_n^{(i)}}\right) = \log w_n^{(i)} - \log\left(\sum_{l=1}^N \exp(\log w_n^{(i)})\right),$$
$$\exp\left(\log W_n^{(i)}\right) = \exp\left(\log\left(\frac{w_n^{(i)}}{\sum_{i=1}^N w_n^{(i)}}\right)\right).$$

Though $\log w_n^{(i)} > 0$, this expression still might suffer from underflow due to the summation term

$\sum_{i=1}^N \exp(\log w_n^{(i)})$. Define $\log w_n^{\min} := \min_i \log w_n^{(i)}$. Then observe that:

$$\begin{aligned}\log\left(\sum_{i=1}^N \exp(\log w_n^{(i)})\right) &= \log\exp(\log w_n^{\max})\sum_{i=1}^N \exp(\log w_n^{(i)} - \log w_n^{\min}),\\ &= \log\left(\sum_{i=1}^N \exp(\log w_n^{(i)} - \log w_n^{\min})\right) + \log w_n^{\min}.\end{aligned}$$

Thus we obtain:

$$\begin{aligned}\log W_n^{(i)} &= \log w_n^{(i)} - \log w_n^{\min} - \log\left(\sum_{i=1}^N \exp(\log w_n^{(i)} - \log w_n^{\min})\right),\\ \iff W_n^{(i)} &= \frac{\exp\left(\log w_n^{(i)} - \log w_n^{\min}\right)}{\sum_{i=1}^N \exp(\log w_n^{(i)} - \log w_n^{\min})}.\end{aligned}$$

This ensures that $\log w_n^{(i)} - \log w_n^{\min} \in [0, \log w_n^{\max} - \log w_n^{\min}]$, and therefore $\exp\left(\log w_n^{(i)} - \log w_n^{\min}\right)$ might not suffer from underflow. Since this simple technique is quite useful in practice, we summarise this as follows.

Algorithm 11 Logsumexp for the normalised weights

i) For $i = 1, \ldots, N$, obtain $\log w_n^{(i)}$.

ii) Calculate $\log w_n^{\min} := \min_i \log w_n^{(i)}$.

iii) For $i = 1, \ldots, N$, obtain $\exp\left(\log w_n^{(i)} - \log w_n^{\min}\right)$.

iv) For $i = 1, \ldots, N$, obtain the normalised weights $W_n^{(i)} = \frac{\exp\left(\log w_n^{(i)} - \log w_n^{\min}\right)}{\sum_{i=1}^N \exp(\log w_n^{(i)} - \log w_n^{\min})}$.

3.3 Sequential Monte Carlo samplers

SMC can be applied to more general settings. That is, SMC can be applied to the targets which are defined *on a common measurable space*. Therefore, one can make use of SMC for static problems for instance. Indeed, *Iterated Batch Importance Sampling* (IBIS) algorithm (Chopin, 2002) and *SMC samplers* (Del Moral et al., 2006) have been widely used for such problem, for instance. Note that both methods are kind of generalisation of *Annealed Importance Sampling* (AIS), studied in Neal (2001). Here we study SMC samplers since this sampler falls under a broader class of SMC methods.

To facilitate the study, we consider the following example. Suppose that now $\pi(dx) \in \mathcal{P}(E)$ is such that:

$$\pi(dx) = \frac{\mathcal{L}(x)\pi_0(dx)}{Z}, \tag{3.15}$$

where $\mathcal{L} : E \to \mathbb{R}$ is a likelihood function, $\pi_0(dx) \in \mathcal{P}(E)$ is a prior distribution, and $Z := \int_E \mathcal{L}(x)\pi_0(dx)$ is a marginal likelihood function so that $\pi(dx)$ corresponds to a posterior distribution. Moreover, sup-

pose that now $(E, \mathcal{E})$ is a high dimensional space. In this case, as noted in Neal (2001), *the tempered posterior:*

$$\pi_n(dx) := \frac{\mathcal{L}(x)^{\phi_n}\pi_0(dx)}{Z_n} \tag{3.16}$$

might provide a beneficial tempering effect and potential reduction in computational complexity, where $\{\phi_n\}_{n=0}^p \subseteq [0,1]$ is an increasing sequence. As a result, now we have the sequence of probability distributions $\{\pi_n(dx)\}_{n=0}^p$ which are defined on a common measurable space. SMC cannot be applied directly to such a sequence of distributions since it is available for distributions whose dimension is increasing over time index as we studied.

To make use of SMC for $\{\pi_n(dx)\}_{n=0}^p$ defined on a common measurable space $(E, \mathcal{E})$, consider the following sequence of auxiliary distributions on the product spaces of increasing dimensions $(E^p, \mathcal{E}^p) := (\prod_{n=0}^p E^n, \mathcal{E}^{\otimes p})$:

$$\mathbb{P}(dx_{0:p}) := \pi_p(dx_p)\prod_{n=1}^{p}\mathcal{B}_{n-1}(x_n, dx_{n-1}), \tag{3.17}$$

where the sequence of artificial backward Markov kernels $\{\mathcal{B}_n\}_{n=0}^{p-1} : E \times \mathcal{E} \to [0,1]$ can be in principle arbitrarily selected. Notice that $\mathbb{P}(dx_{0:p})$ admits marginally $\pi_p(dx_p)$. That is, for any appropriate test function f, we have that $\int f(x_p)\mathbb{P}(dx_{0:p}) = \int f(x_p)\pi_p(dx_p)$. Assume that also we have non-homogeneous Markov kernels $\{\mathcal{K}_n\}_{n=1}^p : E \times \mathcal{E} \to [0,1]$ which are chosen to satisfy that $\mathcal{B}_{n-1} \otimes \gamma_n$ is absolutely continuous w.r.t. $\gamma_{n-1} \otimes \mathcal{K}_n$ for any n. Here, $\gamma_n := \mathcal{L}(x)^{\phi_n}\pi_0(dx)$. In this case, if we define:

$$\mathbb{Q}(dx_{0:p}) := \pi_0(dx_0)\prod_{n=1}^{p}\mathcal{K}_n(x_{n-1}, dx_n), \tag{3.18}$$

then, as a consequence of the Radon–Nikodým theorem, we have that, for an appropriate test function, $\mathbb{E}_{\mathbb{P}}[f(x_p)] = \mathbb{E}_{\mathbb{Q}}\left[f(x_p)\frac{d\mathbb{P}}{d\mathbb{Q}}(x_{0:p})\right]$ with:

$$\frac{d\mathbb{P}}{d\mathbb{Q}}(x_{0:p}) \propto \prod_{n=1}^{p}\frac{\gamma_n(dx_n)\mathcal{B}_{n-1}(x_n, dx_{n-1})}{\gamma_{n-1}(dx_{n-1})\mathcal{K}_n(x_{n-1}, dx_n)}. \tag{3.19}$$

$\mathbb{Q}(dx_{0:p})$ is called the importance distribution, in the literature. Now we can describe a generic SMC algorithm to sample from (3.17). As we noted, generally, SMC iterates sequentially three steps, that is *mutation*, *correction* and *selection*. At the mutation step, we simulate $\{x_n^{(i)}\}_{i=1}^N$ from $\mathcal{K}_n(\tilde{x}_{n-1}^{(i)}, dx_n)$ for $i = 1, 2, \cdots N \in \mathbb{Z}$, and set $x_{0:n}^{(i)} = (x_n^{(i)}, \tilde{x}_{0n-1}^{(i)})$ for each i. As before, these obtained particles are then corrected via the following unnormalised importance weights based on (3.19):

$$w_n^{(i)}(x_{0:n}) := \prod_{k=1}^{n}\frac{\gamma_k(dx_k^{(i)})\mathcal{B}_{k-1}(x_k^{(i)}, dx_{k-1}^{(i)})}{\gamma_{k-1}d(x_{k-1}^{(i)})\mathcal{K}_k(x_{k-1}^{(i)}, dx_k^{(i)})} = w_{n-1}^{(i)}(x_{0:n-1})\frac{\mathcal{B}_{n-1}(x_n^{(i)}, dx_{n-1}^{(i)})}{\mathcal{K}_n(x_{n-1}^{(i)}, dx_n^{(i)})}\frac{\gamma_n(dx_n^{(i)})}{\gamma_{n-1}(dx_{n-1}^{(i)})}. \tag{3.20}$$

It is straightforward to observe that (3.20) becomes:

$$w_n^{(i)}(x_{0:n}) = \prod_{k=1}^{n} G_k(x_{k-1}^{(i)}, x_k^{(i)}), \tag{3.21}$$

$$G_k(x_{k-1}^{(i)}, x_k^{(i)}) := \frac{\mathcal{B}_{k-1}(x_k^{(i)}, dx_{k-1}^{(i)})}{\mathcal{K}_k(x_{k-1}^{(i)}, dx_k^{(i)})} \frac{\gamma_k(dx_k^{(i)})}{\gamma_{k-1}(dx_{k-1}^{(i)})}. \tag{3.22}$$

Given (3.22), next define the normalised weights:

$$W^{(i)}(x_{n-1}^{(i)}, x_n^{(i)}) := \frac{G_n(x_{n-1}^{(i)}, x_n^{(i)})}{\sum_{j=1}^{N} G_n(x_{n-1}^{(j)}, x_n^{(j)})}. \tag{3.23}$$

After the correction step, obtained particles $\{x_{0:n}^{(i)}\}_{i=1}^N$ will be resampled according to their normalised weights $\left\{W_n^{(i)}\right\}_{i=1}^N$ in (3.23), same as Algorithm 9.

Next we consider a choice of backward Markov kernels $\{\mathcal{B}_n\}_{n=0}^{p-1}$. Del Moral et al. (2006, Proposition 1) provide the expression of the sequence of optimal backward Markov kernels $\{\mathcal{B}_n\}_{n=0}^{p-1}$ which minimise the estimator variance, but they are not feasible in general. To approximate the optimal backward Markov kernels, under the implicit assumption that successive targets are similar $\eta_n \approx \eta_{n-1}$ for $n \geq 0$, a common choice of $\{\mathcal{K}_n\}_{n=1}^p$ will be using MCMC kernels which leave $\pi_n(dx_n)$ invariant for any $n \geq 1$. Then, we can construct $\{\mathcal{B}_n\}_{n=0}^{p-1}$ to satisfy:

$$\pi_n(dx_n)\mathcal{B}_{n-1}(x_n, dx_{n-1}) = \pi_n(dx_{n-1})\mathcal{K}_n(x_{n-1}, dx_n), \tag{3.24}$$

for any n. Namely we set:

$$\mathcal{B}_{n-1}(x_n, dx_{p-1}) = \frac{\mathcal{K}_n(x_{n-1}, dx_n)\pi_n(dx_{n-1})}{\pi_n(dx_n)} = \frac{\mathcal{K}_n(x_{n-1}, dx_n)\gamma_n(dx_{n-1})}{\gamma_n(dx_n)}. \tag{3.25}$$

Thus, it turns out that such $\{\mathcal{B}_n\}_{n=0}^{p-1}$ are the time reversal of the $\{\pi_n(dx_n)\}_{n=1}^p$-invariant MCMC kernels $\{\mathcal{K}_n\}_{n=1}^p$. Notice that, with this choice of $\{\mathcal{B}_n\}_{n=0}^{p-1}$, (3.19) now becomes:

$$\frac{d\mathbb{P}}{d\mathbb{Q}}(x_{0:p}) \propto \prod_{n=1}^{p} \frac{\gamma_n(dx_{n-1})}{\gamma_{n-1}(dx_{n-1})} = \prod_{n=1}^{p} G_n(x_{n-1}). \tag{3.26}$$

Notice that when $\pi_n(dx_n)$ is the form of (3.16), under our specification (3.26), it is straightforward to observe that (3.22) becomes:

$$G_n(x_{n-1}^{(i)}) = \frac{\gamma_n(dx_{n-1})}{\gamma_{n-1}(dx_{n-1})} = \mathcal{L}(x_{n-1}^{(i)})^{(\phi_n - \phi_{n-1})}. \tag{3.27}$$

Notice that the expression of (3.26) does not depend on x_n at the current time step n. Therefore, at the time step n, the selection step can be implemented before the mutation step. Indeed, such mutation after correction and selection yields a better approximation of the target since it gives rise to a greater number of distinct particles to approximate the target, see, e.g. Doucet and Johansen

(2009, Section 4.2) and Del Moral et al. (2006, Remark 1). Also, it is clear to see that we do not need to store an entire path $x_{0:n}$. Indeed, only (x_{n-1}, x_n) are needed at the time step n in practice. We summarise the discussed algorithm as follows.

Algorithm 12 SMC samplers (Del Moral et al., 2006)

Assume that at the (auxiliary) time step $n-1$, one has a weighted particles system $\left(x_{n-1}^{(i)}, \frac{1}{N}\right)_{i=1}^{N}$ which approximates $\pi_{n-1}(dx_{n-1})$.

i) For each $i = 1, \ldots, N$, correct the unnormalised importance weights $G_n(x_{n-1}^{(i)})$ via (3.27).

ii) For each $i = 1, \ldots, N$, obtain the normalized importance weights $W_n^{(i)}$ in (3.23).

iii) Do resampling to obtain the equally weighted particles system $\left(\tilde{x}_{n-1}^{(i)}, \frac{1}{N}\right)_{i=1}^{N}$.

iv) For each $i = 1, \ldots, N$, mutate particles to obtain $x_n^{(i)}$ via MCMC kernels $\mathcal{K}_n(\tilde{x}_{n-1}^{(i)}, dx_n)$.

v) Repeat from step1 to step 5 until $n = p$.

Example 12. *SMC sampler for AR(1).*

We simulated AR(1) process in Example 2 with $n = 500$ and $(\phi, \sigma^2) = (0.8, 1.0)$. To carry out SMC sampler for the tempered posterior, we used $1,000$ particles with $5,000$ MCMC iterations. Also, we used random walk Metropolis (Example 10) as MCMC kernel with the proposal density $\mathcal{N}\left(\hat{\theta}_{i-1}; 0, 0.01^2\right)$ at MCMC time step i where $\hat{\theta}_{i-1}$ denotes estimates of the parameters at MCMC time step $i-1$. As for priors, we used $\mathcal{TN}_{[-1,1]}\left(0, 0.5^2\right)$ for ϕ and the inverse gamma distribution with parameters $(3, 3)$ for σ where $\mathcal{TN}_{[a,b]}\left(\mu, \sigma^2\right)$ denotes the truncated Gaussian distribution with mean μ, standard deviation σ in the interval $[a, b]$. To select a schedule of $\{\phi_n\}_{n=0}^{p}$, we applied the adaptive method provided in Beskos et al. (2016). The results are plotted in Figure 1.

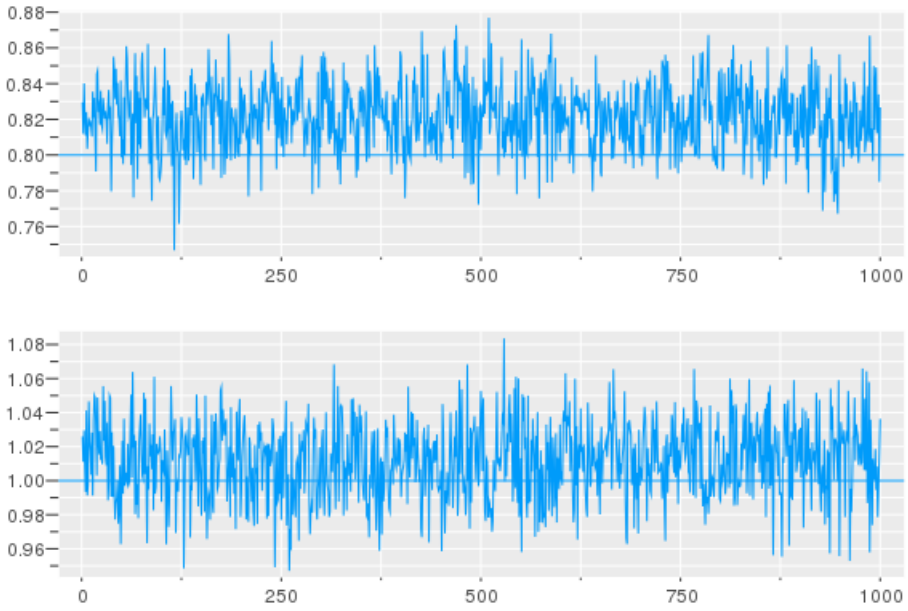


Figure 1: The results of SMC sampler for ϕ (top) and σ^2 (bottom). We used $N = 1,000$ particles with $5,000$ MCMC iterations. The horizontal dash lines indicate the true parameter values in each case.

3.4 Feynman-Kac formulae

Following Del Moral (2004, 2013) closely, this section provides a unified framework to study SMC. As we will see, SMC methods can be understood as a mean field approximation of Feynman-Kac models.

3.4.1 Notations

Let $(E_n, \mathcal{E}_n)$ be a sequence of measurable spaces, and $G_n : E_n \to [0, \infty)$ be a sequence of measurable *potential functions*. Let (x_n) be a non-homogenous Markov chain on a sequence of state-spaces E_n with Markov kernels M_n. For $f_n \in \mathcal{B}_b(E_n)$, $\mu_n \in P(E_n)$, and a Markov kernel M_n, we define the integral $\mu_n(f_n) := \int f_n(x_n)\mu(dx_n)$, the function $M_n(f_n) := \int f_n(x_n)M_n(x_{n-1}, dx_n) \in \mathcal{B}_b(E_n)$ and the probability measure $\mu_n M_{n+1}(dx_{n+1}) := \int \mu_n(dx_n)M_{n+1}(x_n, dx_{n+1}) \in \mathcal{P}(E_n)$. Then for $f_n \in \mathcal{B}_b(E_n)$, we have that $\mathbb{E}[f_n(x_n) \mid x_{n-1}] = \int f_n(x_n)M_n(x_{n-1}, dx_n) = M_n(f_n)$. Also, for $\mu_n \in \mathcal{P}(E_n)$, it can be observed that $\mu_n(dx_n) = \int \mu_{n-1}(dx_{n-1})M_n(x_{n-1}, dx_n) = \mu_{n-1}M_n(dx_n)$. Also we write for $f_n \in \mathcal{B}_b(E_n)$, $\mu_{n-1}M_n(f_n) = \int\int \mu_{n-1}(dx_{n-1})M_n(x_{n-1}, dx_n)f_n(x_n)$.

3.4.2 Basics

Given an initial distribution $\mu_0 \in \mathcal{P}(E_0)$, we write the law of the Markov chain on path space $E^n := \prod_{p=0}^n E_p$, equipped with the product $\mathcal{E}^n := \prod_{p=0}^n \mathcal{E}^{\otimes p}$, as:

$$\mathbb{P}(dx_{0:n}) := \mu_0(dx_0) \prod_{p=1}^{n} M_p(x_{p-1}, dx_p). \tag{3.28}$$

Given (3.28), we can define *the Feynman-Kac models* associated with the pair (G_n, M_n) as follows.

Definition 26. (Feynman-Kac models)

Given the pair (G_n, M_n) and the initial distribution μ, the Feynman-Kac prediction and updated path models are the sequence of path measure defined respectively:

$$\mathbb{Q}(dx_{0:n}) := Z_n^{-1} \prod_{p=0}^{n-1} G_p(x_p)\mathbb{P}(dx_{0:n}), \tag{3.29}$$

$$\hat{\mathbb{Q}}(dx_{0:n}) := \hat{Z}_n^{-1} \prod_{p=0}^{n} G_p(x_p)\mathbb{P}(dx_{0:n}), \tag{3.30}$$

for any $n \in \mathbb{N}$, where

$$Z_n := \mathbb{E}_{\mathbb{P}}\left[\prod_{p=0}^{n-1} G_p(x_p)\right].$$
$$\hat{Z}_n := Z_{n+1}.$$

For further analysis, we introduce the flow of the time marginals of (3.29) and (3.30).

Definition 27. (Time marginals of Feynman-Kac models)

For $f_n \in \mathcal{B}_b(E_n)$, we define the following sequence of positive signed measures:

$$\gamma_n(f_n) := \mathbb{E}_{\mathbb{P}}\left[f_n(x_n)\prod_{p=0}^{n-1} G_p(x_p)\right], \tag{3.31}$$

$$\hat{\gamma}_n(f_n) := \mathbb{E}_{\mathbb{P}}\left[f_n(x_n)\prod_{p=0}^{n} G_p(x_p)\right]. \tag{3.32}$$

(3.31) and (3.32) are known as the unnormalised and updated Feynman-Kac model respectively. We can also define the normalised version of (3.31) and (3.32) as follows, for $f_n \in \mathcal{B}_b(E_n)$,

$$\eta_n(f_n) := \gamma_n(f_n)/\gamma_n(1),\ \hat{\eta}_n(f_n) := \hat{\gamma}_n(f_n)/\hat{\gamma}_n(1). \tag{3.33}$$

Notice that $\eta_n \in \mathcal{P}(E_n)$. First, observe that:

$$\eta_n(G_n) = \frac{\gamma_n(G_n)}{\gamma_n(1)} = \frac{\gamma_{n+1}(1)}{\gamma_n(1)},$$
$$\iff \gamma_{n+1}(1) = \eta_n(G_n) \times \gamma_n(1),$$

and this implies:

$$\gamma_n(1) = \eta_{n-1}(G_{n-1})\gamma_{n-1}(1),$$
$$= \eta_{n-1}(G_{n-1})\eta_{n-2}(G_{n-2})\gamma_{n-2}(1),$$

and so on. Thus we have:

$$\gamma_{n+1}(1) = \prod_{0\leq p\leq n} \eta_p(G_p), \tag{3.34}$$

where $\gamma_{n+1}(1) = Z_{n+1} = \hat{Z}_n$. Next, for any $f_n \in \mathcal{B}_b(E_n)$, we have that:

$$\gamma_n(f_n \times G_n) = \mathbb{E}_{\mathbb{P}}\left[f_n(x_n)G_n(x_n)\prod_{p=0}^{n-1} G_p(x_p)\right] = \hat{\gamma}_n(f_n). \tag{3.35}$$

Therefore, by the definition in (3.33), we have that:

$$\hat{\eta}_n(f_n) = \frac{\gamma_n(f_n \times G_n)}{\gamma_n(G_n)} = \frac{\gamma_n(f_n \times G_n)/\gamma_n(1)}{\gamma_n(G_n)\gamma_n(1)} = \frac{\eta_n(f_n \times G_n)}{\eta_n(G_n)}.$$

This observation leads to the following the *Boltzmann-Gibbs transformation.*

Definition 28. *The Boltzmann-Gibbs transformation is the mapping $\Psi_{G_n} : \mathcal{P}(E_n) \to \mathcal{P}(E_n)$, defined for all $\mu_n \in \mathcal{P}(E_n)$ by the following measure:*

$$\Psi_{G_n}(\mu_n)(dx_n) := \frac{G_n(x_n)\mu_n(dx_n)}{\mu_n(G_n)}. \tag{3.36}$$

Using (3.36), we can re-write $\hat{\eta}_n$ as follows, for any $f_n \in \mathcal{B}_b(E_n)$:

$$\hat{\eta}_n(f_n) = \Psi_{G_n}(\eta_n(f_n)). \tag{3.37}$$

Next, due to the Markovian property, for any $f_n \in \mathcal{B}_b(E_n)$, we have that:

$$\begin{aligned}\gamma_n(f_n) &= \mathbb{E}_{\mathbb{P}}\left[f_n(x_n)\prod_{p=0}^{n-1} G_p(x_p)\right] = \mathbb{E}_{\mathbb{P}}\left[\mathbb{E}_{\mathbb{P}}\left[f_n(x_n)\prod_{p=0}^{n-1} G_p(x_p) \mid x_{0:n-1}\right]\right] \\ &= \mathbb{E}_{\mathbb{P}}\left[\mathbb{E}_{\mathbb{P}}\left[f(x_n) \mid x_{n-1}\right]\prod_{p=0}^{n-1} G_p(x_p)\right] = \mathbb{E}_{\mathbb{P}}\left[M_n(f_n) \times \prod_{p=0}^{n-1} G_p(x_p)\right].\end{aligned}$$

That is, we have that:

$$\gamma_n(f_n) = \hat{\gamma}_{n-1}M_n(f_n). \tag{3.38}$$

Recall that $\eta_n(f_n) := \gamma_n(f_n)/\gamma_n(1)$. So, (3.37) and (3.38) give rise to:

$$\eta_n(f_n) = \frac{\hat{\gamma}_{n-1}M_n(f_n)}{\hat{\gamma}_{n-1}(1)} = \hat{\eta}_{n-1}M_n(f_n). \tag{3.39}$$

We are now ready to derive the following basic recursions.

Proposition 23. *For any $f_n \in \mathcal{B}_b(E_n)$, we have that:*

$$\gamma_n(f_n) = \eta_n(f_n) \prod_{p=0}^{n-1} \eta_p(G_p),$$
$$\hat{\gamma}_n(f_n) = \hat{\eta}_n(f_n) \prod_{p=0}^{n} \eta_p(G_p),$$
$$\eta_n(f_n) = \Psi_{G_{n-1}}(\eta_{n-1}) M_n(f_n),$$
$$\hat{\eta}_n(f_n) = \Psi_{G_n}(\hat{\eta}_{n-1} M_n(f_n)).$$

Proof. $\eta_n(f_n)\gamma_n(1) = \gamma_n(f_n)$ and $\gamma_{n+1}(1) = \prod_{0 \leq p \leq n} \eta_p(G_p)$ by (3.34), thus we have $\gamma_n(f_n) = \eta_n(f_n) \prod_{p=0}^{n-1} \eta_p(G_p)$. Also, notice that $\hat{\eta}_n(f_n)\hat{\gamma}_n(1) = \hat{\gamma}_n(f_n)$ and $\hat{\gamma}_n(1) = \gamma_{n+1}(1)$, thus $\hat{\gamma}_n(f_n) = \hat{\eta}_n(f_n) \prod_{p=0}^{n} \eta_p(G_p)$. From (3.39) and (3.37), we have that $\eta_n(f_n) = \hat{\eta}_{n-1} M_n(f_n) = \Psi_{G_{n-1}}(\eta_{n-1}) M_n(f_n)$. In the same manner, we have that $\hat{\eta}_n(f_n) = \Psi_{G_n}(\eta_n(f_n)) = \Psi_{G_n}(\hat{\eta}_{n-1} M_n(f_n))$. □

3.4.3 Feynman-Kac semigroup models

Recall that for any $f_n \in \mathcal{B}_b(E_n)$, $\gamma_n(f_n) = \hat{\gamma}_{n-1} M_n(f_n)$ holds. Also recall that, for any $f_n \in \mathcal{B}_b(E_n)$, $\hat{\gamma}_n(f_n) = \gamma_n(G_n \times f_n)$. Then, for any n, define the kernel Q_{n+1} from $(E_n, \mathcal{E}_n)$ into $(E_{n+1}, \mathcal{E}_{n+1})$ such that, given for any $x_n \in E_n$ by:

$$Q_{n+1}(x_n, dx_{n+1}) := G_n(x_n) \times M_n(x_n, dx_{n+1}). \tag{3.40}$$

Clearly, we can show that for any $f_n \in \mathcal{B}_n(E_n)$:

$$\gamma_n(f_n) = \gamma_{n-1} Q_n(f_n),$$

holds. Therefore, we have proven that, for any $f_n \in \mathcal{B}_b(E_n)$:

$$\begin{aligned} \gamma_n(f_n) &= \gamma_{n-1} Q_n(f_n) = \gamma_{n-2} Q_{n-1} Q_n(f_n) = \cdots \\ &= \gamma_p Q_{p:n}(f_n), \end{aligned} \tag{3.41}$$

where the linear semigroup is defined as $Q_{p:n} := Q_{p+1} \cdots Q_n$ for $p < n$ with $Q_{n:n} = I$, here I is the identity operator. Notice that for $f_n \in \mathcal{B}_b(E_n)$,

$$Q_{p:n}(f_n) = \mathbb{E}_{\mathbb{P}(x_{p:n})} \left[f(x_n) \prod_{q=p}^{n-1} G_q(x_q) \right].$$

Next, from Proposition 23, we know that $\eta_n(f_n) = \Psi_{G_n}(\eta_{n-1}) M_n(f_n)$, and define the mapping $\Phi_n : \mathcal{P}(E_{n-1}) \longmapsto \mathcal{P}(E_n)$:

$$\Phi_n(\eta_{n-1}) := \Psi_{G_n}(\eta_{n-1}) M_n, \tag{3.42}$$

that is, we have that, for $f_n \in \mathcal{B}_n(E_n)$:

$$\eta_n(f_n) = \Phi_n(\eta_{n-1})(f_n). \tag{3.43}$$

Given this, we can obtain the nonlinear semigroup associated to the normalised Feynman-Kac measures $\{\eta_n\}$ such that:

$$\Phi_{p:n} := \Phi_n \circ \Phi_{n-1} \circ \cdots \circ \Phi_{p+1}, \tag{3.44}$$

for any $0 \leq p \leq n$ with $\Phi_{n:n} = I$. Using the semigroup, we can write:

$$\eta_n(f_n) = \Phi_{p:n}(\eta_p)(f_n).$$

Indeed, from the definition of η_n, Proposition 23, (3.40) and (3.42), we can show that, for any $f_n \in \mathcal{B}_n(E_n)$:

$$\begin{aligned} \eta_n(f_n) &:= \frac{\gamma_n(f_n)}{\gamma_n(1)} = \frac{\gamma_p Q_{p:n}(f_n)}{\gamma_p Q_{p:n}(1)} = \frac{\gamma_p Q_{p:n}(f_n)/\gamma_p(1)}{\gamma_p Q_{p:n}(1)/\gamma_p(1)} = \frac{\eta_p Q_{p:n}(f_n)}{\eta_p Q_{p:n}(1)}, \\ &= \Phi_{p:n}(\eta_p)(f_n). \end{aligned} \tag{3.45}$$

All in all, now we are ready to state the following proposition.

Proposition 24. *For any $f_n \in \mathcal{B}_b(E_n)$ and $0 \leq p \leq n$, we have that:*

$$\Phi_{p:n}(\eta_p)(f_n) = \frac{\eta_p Q_{p:n}(f_n)}{\eta_p Q_{p:n}(1)}.$$

We end this section by noting that this nonlinear semigroup is also associated to:

$$P_{p:n} := \frac{Q_{p:n}}{\eta_p Q_{p:n}(1)} = \frac{\gamma_p(1)}{\gamma_n} Q_{p:n}, \tag{3.46}$$

and this implies $\eta_n = \eta_p P_{p:n}$.

3.4.4 Change of measures

This subsection gives a brief explanation of some change of measure techniques used in SMC methods. Assume the one has another collection of Markov kernels $\bar{M}_n(x_{n-1}, dx_n)$ from E_{n-1} into E_n such that for any $x_{n-1} \in E_{n-1}$, $M_n(x_{n-1}, dx_n)$ is absolutely continuous w.r.t. $\bar{M}_n(x_{n-1}, dx_n)$. Also assume that corresponding initial distributions μ_0 and $\bar{\mu}_0$ are so. Define the law of the Markov chain on a path space $E^n := \prod_{p=0}^n E_p$, equipped with the product $\mathcal{E}^n := \prod_{p=0}^n \mathcal{E}_p$, as:

$$\bar{\mathbb{P}}(dx_{0:n}) := \bar{\mu}_0(dx_0) \prod_{p=1}^n \bar{M}_p(x_{p-1}, dx_p), \tag{3.47}$$

Since $\mathbb{P} \ll \bar{\mathbb{P}}$ by construction, via the Radon–Nikodym theorem, we have that:

$$\frac{d\mathbb{P}}{d\bar{\mathbb{P}}}(x_{0:n}) = \frac{\mu_0(dx_0)}{\bar{\mu}_0(dx_0)} \prod_{p=1}^{n} \frac{dM_p(x_{p-1}, \cdot)}{d\bar{M}_p(x_{p-1}, \cdot)}(x_p). \tag{3.48}$$

Let $\mathbb{Q}$ be the Feynman-Kac path measure induced by $\mathbb{P}$. Given this, we can obtain the following positive bounded functions for any $p \geq 1$:

$$\bar{G}_p(x_{p-1}, x_p) := \frac{\mathbb{Q}(dx_{0:p})}{\mathbb{Q}(dx_{0:p-1}) \times \bar{M}_p(x_{p-1}, dx_p)} = \frac{dQ_p(x_{p-1}, \cdot)}{d\bar{M}_p(x_{p-1}, \cdot)}(x_p), \tag{3.49}$$

with $\bar{G}_0 := G_0(x_0)\frac{\mu_0(dx_0)}{\bar{\mu}_0(dx_0)}$, where $Q_p(x_{p-1}, dx_p)$ is given in (3.40). Given the pair $(\bar{G}_n, \bar{M}_n)$ and the initial distribution $\bar{\mu}$, we can define the new Feynman-Kac path measure:

$$\begin{aligned} \bar{\mathbb{Q}}(dx_{0:n}) &:= \bar{Z}_n^{-1} \prod_{p=0}^{n-1} \bar{G}_p(x_{p-1}, x_p) \bar{\mathbb{P}}(dx_{0:n}), \\ \bar{Z}_n &:= \mathbb{E}_{\bar{\mathbb{P}}} \left[\prod_{p=0}^{n-1} \bar{G}_p(x_{p-1}, x_p) \right]. \end{aligned} \tag{3.50}$$

Critically, for any $f_n \in \mathcal{B}_b(E^n)$, it can be shown that:

$$\begin{aligned} \bar{\mathbb{Q}}(f_n) &= \bar{Z}_n^{-1} \int f_n(x_{0:n}) \prod_{p=0}^{n-1} \bar{G}_p(x_{p-1}, x_p) \prod_{p=0}^{n} \bar{M}_p(x_{p-1}, dx_p), \\ &= Z_n^{-1} \int f_n(x_{0:n}) \prod_{p=0}^{n-1} G_p(x_p) \prod_{p=0}^{n} M_p(x_{p-1}, dx_p), \\ &= \mathbb{Q}(f_n), \end{aligned}$$

and the same thing holds for updated path model $\hat{\mathbb{Q}}(dx_{0:n})$. From these observations, we have proven the following preposition.

Proposition 25. *The Feynman-Kac prediction and updated path models associated (G_n, M_n) and $(\bar{G}_n, \bar{M}_n)$ coincide. Also, their time marginals coincide.*

Example 13. *SMC samplers.*

Let $\pi(dx) = Z^{-1}\bar{\pi}(dx) \in \mathcal{P}(E)$ be a target distribution from which one wishes to sample. Consider the following tempered distributions:

$$\begin{aligned} \pi_n(dx) &:= \bar{\pi}_n(dx) Z_n^{-1}, \\ \bar{\pi}_n(dx) &:= \mu(dx) \left(\frac{d\bar{\pi}}{d\mu}(x) \right)^{\phi_n}, \end{aligned}$$

where $\pi \ll \mu \in \mathcal{P}(E)$, $Z_n = \int \bar{\pi}_n(x_n) dx_n$ and $0 = \phi_0 < \phi_1 < \cdots < \phi_p = 1$. Clearly, $\pi_p(dx) = \pi(dx)$ and $\pi_0(dx) = \mu(dx)$. Then one has a sequence of distributions $\{\pi_n(dx)\}_{n=0}^{p} \in \mathcal{P}(E)$ defined on the same the measurable space $(E, \mathcal{E})$, thus SMC cannot be directly applied. Consider the following

artificial extended target distribution on $(E^p, \mathcal{E}^p)$:

$$\mathbb{Q}(dx_{0:p}) = \pi_p(dx_p) \prod_{n=1}^{p} \mathcal{B}_{n-1}(x_n, dx_{n-1}),$$

where $\{\mathcal{B}_n\}$ is a is a sequence of Markov kernels from E_n into E_{n-1}. Also, let $\{\mathcal{K}_n\}$ be a collections of Markov chains from E_{n-1} into E_n and assume that, for all n, $\bar{\pi}_n(dx_n)\mathcal{B}_{n-1}(x_n, dx_{n-1}) \ll \bar{\pi}_{n-1}(dx_{n-1})\mathcal{K}_n(x_{n-1}, dx_n)$ holds. As before, we define:

$$\bar{\mathbb{P}}(dx_{0:p}) = \pi_0(dx_0) \prod_{n=1}^{p} \mathcal{K}_n(x_{n-1}, dx_n).$$

Moreover, we define the appropriate potential function $\bar{G}_n(x_{n-1}, x_n) : E \times E \to [0, \infty)$ such as:

$$\bar{G}_n(x_{n-1}, x_n) = \frac{\bar{\pi}_n(dx_n)\mathcal{B}_{n-1}(x_n, dx_{n-1})}{\bar{\pi}_{n-1}(dx_{n-1})\mathcal{K}_n(x_{n-1}, dx_n)},$$

for any $n \geq 1$ with $\bar{G}_0(x_0) = 1$. If we admit the corresponding densities, we have that:

$$\bar{G}_n(x_{n-1}, x_n) = \frac{\bar{\pi}_n(x_n)\mathcal{B}_{n-1}(x_n, x_{n-1})}{\bar{\pi}_{n-1}(x_{n-1})\mathcal{K}_n(x_{n-1}, x_n)}.$$

With these ingredients, set:

$$\bar{\mathbb{Q}}(dx_{0:p}) = \bar{Z}_p^{-1} \prod_{n=0}^{p} \bar{G}_n(x_{n-1}, x_n)\bar{\mathbb{P}}(dx_{0:p}),$$
$$\bar{Z}_n := \mathbb{E}_{\bar{\mathbb{P}}}\left[\prod_{n=0}^{p} \bar{G}_n(x_{n-1}, x_n)\right].$$

Then, for any $f_n \in \mathcal{B}_b(E^n)$, observe that:

$$\begin{aligned}
\bar{\mathbb{Q}}(f_n) &= \bar{Z}_n^{-1} \int f_n(x_{0:n})\bar{G}_0(x_0) \prod_{p=1}^{n} \frac{\bar{\pi}_p(x_p)\mathcal{L}_{p-1}(x_p, x_{p-1})}{\bar{\pi}_{p-1}(x_{p-1})\mathcal{K}_p(x_{p-1}, x_p)} \pi_0(x_0)\mathcal{K}_p(x_{p-1}, x_p)dx_{0:p}, \\
&= \int f_n(x_{0:n})\pi_n(x_n) \prod_{p=1}^{n} \mathcal{L}_{p-1}(x_p, x_{p-1})dx_{0:p}, \\
&= \mathbb{Q}(f_n),
\end{aligned}$$

thus the pair $(\bar{G}_n(x_{n-1}, x_n), \mathcal{K}_n)$ recovers the Feynman-Kac path measure $\mathbb{Q}(dx_{0:n})$, and thus SMC

can be applied. As for time marginals, for any $f_n \in \mathcal{B}_b(E_n)$, it can be shown that:

$$\begin{aligned}
\hat{\gamma}_n(f_n) &= \mathbb{E}_{\bar{\mathbb{P}}}\left[f_n(x_n)\prod_{p=0}^{n}\bar{G}_p(x_{p-1},x_p)\right], \\
&= \int f_n(x_n)\bar{G}_0\prod_{p=1}^{n}\frac{\bar{\pi}_p(x_p)\mathcal{L}_{p-1}(x_p,x_{p-1})}{\bar{\pi}_{p-1}(x_{p-1})\mathcal{K}_p(x_{p-1},x_p)}\pi_0(x_0)\mathcal{K}_p(x_{p-1},x_p)dx_{0:p}, \\
&= \int f_n(x_n)\bar{\pi}_n(x_n)\prod_{p=1}^{n}\mathcal{L}_{p-1}(x_p,x_{p-1})dx_{0:p}, \\
&= \int f_n(x_n)\bar{\pi}_n(x_n)dx_n,
\end{aligned}$$

and $\hat{\gamma}_n(1) = \int \bar{\pi}_n(x_n)dx_n = Z_n$, and thus $\hat{\eta}_n(f_n) = \int f_n(x_n)\pi_n(x_n)dx_n$ in the case of SMC samplers. That is, time marginal (updated) normalised Feynman-Kac models in (3.33) act as the sequence of the tempered distributions $\{\pi_n(dx)\}$. Assume that $\{\mathcal{K}_n\}$ are MCMC kernels which leave $\pi_n(x)$ invariant, and one can set:

$$\mathcal{L}_{n-1}(x_n,x_{n-1}) = \frac{\pi_n(x_{n-1})\mathcal{K}_n(x_{n-1},x_n)}{\pi_n(x_n)},$$

that is, $\{\mathcal{L}_n\}$ are time reversal Markov kernels in the sense:

$$\pi_n(dx_n)\mathcal{L}_{n-1}(x_n,dx_{n-1}) = \pi_n(dx_{n-1})\mathcal{K}_n(x_{n-1},dx_n).$$

In this case, we have that:

$$\begin{aligned}
\bar{G}_n(x_{n-1},x_n) &= \frac{\bar{\pi}_n(x_n)\mathcal{L}_{n-1}(x_n,x_{n-1})}{\bar{\pi}_{n-1}(x_{n-1})\mathcal{K}_n(x_{n-1},x_n)} = \frac{\bar{\pi}_n(x_n)\pi_n(x_{n-1})\mathcal{K}_n(x_{n-1},x_n)}{\bar{\pi}_{n-1}(x_{n-1})\pi_n(x_n)\mathcal{K}_n(x_{n-1},x_n)}, \\
&= \frac{\bar{\pi}_n(x_n)\bar{\pi}_n(x_{n-1})/Z_n}{\bar{\pi}_{n-1}(x_{n-1})\bar{\pi}_n(x_n)/Z_n} = \frac{\bar{\pi}_n(x_{n-1})}{\bar{\pi}_{n-1}(x_{n-1})}.
\end{aligned}$$

3.5 Mean field interacting particle models

3.5.1 Mckean interpretation

In this section we design a non-linear Markov interpretation of the flow of Feynman-Kac models with associated with a pair of (G_n, M_n). Without loss of generality, we assume that for any n, $\|G_n\|_\infty < \infty$. First we define the following new Markov kernel from $(E_n, \mathcal{E}_n)$ into $(E_n, \mathcal{E}_n)$ itself as follows:

$$S_{n,\eta}(x_n, dy_n) := \epsilon_n G_n(x_n)\delta_{x_n}(dy_n) + (1 - \epsilon_n G_n(x_n))\Psi_{G_n}(\eta)(dy_n), \tag{3.51}$$

where ϵ_n is a non-negative constant such that $\epsilon_n G_n(x_n) \leq 1$ for all $x_n \in E_n$. Clearly, we can show that, for any $f_n \in \mathcal{B}_b(E_n)$:

$$S_{n,\eta}(f_n) = \epsilon_n G_n(x_n) \times f_n(x_n) + (1 - \epsilon_n G_n(x_n))\Psi_{G_n}(\eta_n)(f_n),$$

from which we have that:

$$\begin{aligned}\eta_n S_{n,\eta_n}(f_n) &= \epsilon\eta_n(G_n(x_n)\times f_n(x_n)) + (1-\epsilon_n\eta_n(G_n(x_n)))\Psi_{G_n}(\eta_n)(f_n),\\ &= \epsilon\eta_n(G_n(x_n)\times f_n(x_n)) + \Psi_{G_n}(\eta_n)(f_n) - \epsilon_n\eta_n(G_n(x_n))\frac{\eta_n(f_n\times G_n(x_n))}{\eta_n(G_n(x_n))},\\ &= \epsilon\eta_n(G_n(x_n)\times f_n(x_n)) + \Psi_{G_n}(\eta_n)(f_n) - \epsilon_n\eta_n(f_n\times G_n(x_n)) = \Psi_{G_n}(\eta_n)(f_n).\end{aligned}$$

Therefore, $S_{n,\eta}(x_n, dy_n)$ can be considered as an alternative interpretation of $\Psi_{G_n}(\eta)$. Using (3.51), we have that, for any $f_{n+1}\in\mathcal{B}_b(E_{n+1})$:

$$\eta_{n+1}(f_{n+1}) = \eta_n K_{n+1,\eta_n}(f_{n+1}), \tag{3.52}$$

where $K_{n+1,\eta}$ is a collection of Markov kernels from $(E_n,\mathcal{E}_n)$ into $(E_{n+1},\mathcal{E}_{n+1})$ such that:

$$K_{n+1,\eta} := S_{n,\eta}M_{n+1}, \tag{3.53}$$

(3.52) is true since $\eta_{n+1}(f_{n+1}) = \Psi_{G_n}(\eta_n)M_{n+1}(f_{n+1})$ holds (Proposition 23), $\eta_n S_{n,\eta_n} = \Psi_{G_n}(\eta_n)$ and thus $\eta_{n+1}(f_{n+1}) = S_{n,\eta}M_{n+1}(f_{n+1})$. Notice that (3.53) can be decomposed into two separate transitions:

$$\eta_n \overset{S_{n,\eta_n}}{\longrightarrow} \hat{\eta}_n = \eta_n S_{n,\eta_n} = \Psi_{G_n}(\eta_n) \overset{M_{n+1}}{\longrightarrow} \eta_{n+1} = \hat{\eta}_n M_{n+1} = \Psi_{G_n}(\eta_n)M_{n+1} \tag{3.54}$$

Definition 29. (The Mckean measure)

The Mckeam measure associated with a collection of the Mckean-Markov kernels $(K_{n+1,\eta})_{\eta\in\mathcal{P}(E_n)}$ *with the initial distribution* $\eta_0\in\mathcal{P}(E_0)$ *is a probability measure such that:*

$$\mathbb{K}_n(dx_{0:n}) := \eta_0(dx_0)\prod_{p=1}^{n} K_{n+1,\eta_n}(x_{n-1}, dx_n), \tag{3.55}$$

where $\eta_n\in\mathcal{P}(E_n)$ *is the solution of the equation:*

$$\eta_{n+1} = \eta_n K_{n+1,\eta_n}$$

Let $(\bar{x}_n)$ be a Markov chain on E_n with initial distribution $\eta_0 = Law(X_0)$, and elementary Markov kernels given by:

$$\mathbb{P}(\bar{x}_n\in dx \mid \bar{x}_{n-1}) = K_{n+1,\eta_n}(\bar{x}_{n-1}, dx),$$

with $Law(\bar{x}_n) = \eta_n$. Then any $f_n\in\mathcal{B}_b(E_n)$, we have that:

$$\begin{aligned}\mathbb{E}_{\mathbb{K}}\left[f_n(\bar{x}_n)\right] &= \int f_n(\bar{x}_n)\eta_0(dx_0)\prod_{p=1}^{n} K_{n+1,\eta_n}(\bar{x}_{n-1}, dx_n),\\ &= \int f_n(\bar{x}_n)\eta_{n-1}K_{n+1,\eta_n}(\bar{x}_{n-1}, dx_n) = \eta_n(f_n),\end{aligned}$$

thus we conclude that η_n is the law of $\bar{x}_n$ under $\mathbb{K}$. Recall that $\eta_{n+1} = \Phi_{n+1}(\eta_n) := \Psi_{G_{n+1}}(\eta_n)M_{n+1}$, and thus $\Phi_{n+1}(\eta_n) = \eta_n K_{n+1,\eta_n}$. In the literature, such interpretation of Feynman-Kac models is called *Mckean interpretation*. Under $\mathbb{K}$, the motion of $\{\bar{x}_n\}$ can be understood as follows. Given the position and the distribution of $\bar{x}_n$ at time n, w.p. $\epsilon_n G_n(\bar{x}_n)$, $\bar{x}_n$ remains the same position, and we can set $\tilde{x}_n = \bar{x}_n$. In the same manner, w.p. $(1 - \epsilon_n G_n(x_n))$, $\bar{x}_n$ *jumps to the new position which is randomly selected according to the Boltzmann-Gibbs transformation* $\tilde{x}_n = \Psi_{G_n}(\eta_n)(dx_n)$, and we can also set $\tilde{x}_n = \bar{x}_n$. *Then,* $\bar{x}_n$ *evolves to a new site* $\bar{x}_{n+1}$ *via* $M_{n+1}(\tilde{x}_n, \cdot)$. Notice if one sets $\epsilon_n = 0$ for any n, then $\{\bar{x}_n\}$ always jump to a new place, since $S_{n,\eta}(x_n, dy_n) = \Psi_{G_n}(\eta)(dy_n)$ in this case.

3.5.2 Interacting particle systems

In general, we can rarely compute explicitly the law η_n of interest, and this motivates us to consider how to approximate η_n and associated Feynman-Kac models. We adopt Mckean interpretation of Feynman-Kac models. Let N be a positive integer, and define $(E_n^N, \mathcal{E}_n^N) := \left(\prod_{p=1}^N E_n^p, \prod_{p=1}^N \mathcal{E}_n^p\right)$. Then given Mckeam-Markov kernels $(K_{n+1,\eta})_{\eta \in \mathcal{P}(E_n)}$ and the initial distribution $\eta_0 \in \mathcal{P}(E_0^N)$, define a sequence of nonhomologous Markov chains taking values at each time $n \in \mathbb{N}$ in the product space E_n^N such that:

$$\xi_n := (\xi_n^{(1)}, \cdots \xi_n^{(N)})_{n \geq 0},$$

with the elementary transitions from E_{n-1}^N into E_n^N defined as:

$$\mathbb{P}(\xi_n \in \mathbf{x}_n \mid \xi_{n-1}) := \prod_{p=1}^N K_{n,\eta_{n-1}^N(\xi_{n-1})}(\xi_{n-1}, d\mathbf{x}_n), \tag{3.56}$$

where we have defined $\mathbf{x}_n := (x_n^1, \cdots x_n^N)$ and $\eta_n^N(\xi_n) := \frac{1}{N}\sum_{i=1}^N \delta_{\xi_n^{(i)}}(d\xi_n)$. Replacing η_n with its empirical measure, we find that:

$$\begin{aligned} K_{n+1,\eta_n^N} &:= S_{n,\eta_n^N} M_{n+1}, \\ S_{n,\eta_n^N}(\xi_n^{(i)}, \cdot) &:= \epsilon_n G_n(\xi_n^{(i)})\delta_{\xi_n^{(i)}} + (1 - \epsilon_n G_n(\xi_n^{(i)}))\Psi_{G_n}(\eta_n^N(\xi_n)), \end{aligned} \tag{3.57}$$

where the corresponding Boltzmann-Gibbs transformation is given by:

$$\Psi_{G_n}(\eta_n^N(\xi_n)) := \sum_{i=1}^N \frac{G_n(\xi_n^{(i)})}{\sum_{j=1}^N G_n(\xi_n^{(j)})}\delta_{\xi_n^{(i)}}. \tag{3.58}$$

If we set $\epsilon_n = 0$ for any $n \in \mathbb{N}$, we can alternatively write (3.56) as:

$$\mathcal{K}_n(\xi_{n-1}, d\xi_n) := \prod_{p=1}^N \sum_{i=1}^N \frac{G_{n-1}(\xi_{n-1}^{(i)})}{\sum_{k=1}^N G_{n-1}(\xi_{n-1}^{(k)})} M_n(\xi_{n-1}^{(i)}, d\xi_n^{(p)}). \tag{3.59}$$

We call $\{\xi_n\}_{n\geq 0} \in E_n^N$ *particles*. Then (3.54) can be replaced with the evolution of $\{\xi_n\}$ as follows:

$$\xi_n \in E_n^N \overset{selection}{\longrightarrow} \hat{\xi}_n \in E_n^N \overset{mutation}{\longrightarrow} \xi_{n+1} \in E_{n+1}^N.$$

At the *selection* stage, only depends of the potential function G_n, each particle $\xi_n^{(i)}$ is remained at the same position w.p. $\epsilon_n G_n(\xi_n^{(i)})$, and we set $\hat{\xi}_n^{(i)} = \xi_n^{(i)}$. Or, w.p. $\epsilon_n G_n(\xi_n^{(i)})$, we select randomly $\tilde{\xi}_n^{(i)}$ with distribution $\sum_{i=1}^N \frac{G_n(\xi_n^{(i)})}{\sum_{j=1}^N G_n(\xi_n^{(j)})}\delta_{\xi_n^{(i)}}$, and set $\hat{\xi}_n^{(i)} = \tilde{\xi}_n^{(i)}$. Notice that the selection stage depends on only ϵ_n and the current potential function G_n. Given $\{\hat{\xi}_n^{(i)}\}_{i=1}^N$, each selected particle $\hat{\xi}_n^{(i)}$ evolves independently, randomly according to the Markov kernel $M_{n+1}(\hat{\xi}_n^{(i)}, \cdot)$. This stage is called *mutation* in the literature, and only depends on the Markov kernel M_{n+1}. All in all, the genetic type evolution of the interacting particle systems is summarised by the following:

$$\begin{bmatrix} \xi_n^{(1)} \\ \vdots \\ \xi_n^{(i)} \\ \vdots \\ \xi_n^{(N)} \end{bmatrix} \in E_n^N \overset{S_{n,\eta_n^N}(\xi_n^{(i)},\cdot)}{\longrightarrow} \begin{bmatrix} \hat{\xi}_n^{(1)} & \overset{M_{n+1}}{\longrightarrow} & \xi_{n+1}^{(1)} \\ \vdots & & \vdots \\ \hat{\xi}_n^{(i)} & \overset{M_{n+1}}{\longrightarrow} & \xi_{n+1}^{(i)} \\ \vdots & & \vdots \\ \hat{\xi}_n^{(N)} & \overset{M_{n+1}}{\longrightarrow} & \xi_{n+1}^{(N)} \end{bmatrix} \in E_{n+1}^N$$

In the further study of this paper, we are mainly concerned with the convergence analysis of the $n-$time marginal measures η_n^N. From Proposition 23, we can develop the following mean field approximations of time marginals of Feynman-Kac models as follows, for any $f_n \in \mathcal{B}_b(E_n)$:

$$\eta_n^N(f_n) := \frac{1}{N}\sum_{i=1}^N f_n(\xi_n^{(i)}), \tag{3.60}$$

$$\gamma_n^N(f_n) := \eta_n^N(f_n) \times \prod_{p=0}^{n-1} \eta_n^N(G_n), \tag{3.61}$$

note that (3.61) follows from $\gamma_n(1) = \prod_{0\leq p\leq n-1} \eta_p(G_p)$ in (3.34) and the definition $\eta_n(f_n) := \gamma_n(f_n)/\gamma_n(1)$. As in (3.52), we can write:

$$\eta_{n+1}^N = \eta_n^N K_{n+1,\eta_n^N}. \tag{3.62}$$

Then, flow of $\{\eta_n^N\}$ can be also described as:

$$\eta_n^N \overset{selection/resampling}{\longrightarrow} \hat{\eta}_n^N = \eta_n^N S_{n,\eta_n^N} = \Psi_{G_n}(\eta_n^N) \overset{mutaiton}{\longrightarrow} \eta_{n+1}^N = \hat{\eta}_n^N M_{n+1} = \Psi_{G_n}(\eta_n^N) M_{n+1} \tag{3.63}$$

and we algorithmically summarise the discussion. We note that the following algorithm is essentially same as SMC (Algorithm 9).

Algorithm 13 Interacting mean field approximation of the time marginals of Feynman-Kac models

i) *Initialization*: At time $n = 0$, sample N independent random variables ξ_0 from $\eta_0 \in \mathcal{P}(E_0^N)$.

ii) *Selection:* At time $n \geq 1$, given $\left\{\xi_n^{(i)}\right\}_{i=1}^N$, set $\hat{\xi}_n^{(i)} = \xi_n^{(i)}$ w.p. $\epsilon_n G_n(\xi_n^{(i)})$ for each i. Otherwise, select randomly $\tilde{\xi}_n^{(i)}$ with distribution $\sum_{i=1}^N \frac{G_n(\xi_n^{(i)})}{\sum_{j=1}^N G_n(\xi_n^{(j)})}\delta_{\xi_n^{(i)}}$, and set $\hat{\xi}_n^{(i)} = \tilde{\xi}_n^{(i)}$ for each i.

iii) *Mutation:* At time $n \geq 1$, given $\left\{\hat{\xi}_n^{(i)}\right\}_{i=1}^N$, sample conditionally independently $\xi_{n+1}^{(i)}$ from the Markov kernel for each i.

iv) Repeat selection step and mutation step.

3.6 Analysis

This section provides the convergence results of the interacting particle systems (Algorithm 13) following closely Del Moral (2004, 2013); Vergé et al. (2015). To do so, without loss of generality, we assume that $\epsilon_n = 0$ for any n in (Algorithm 13) throughout the rest of the section. Again, this means that we do resampling every time.

3.6.1 Unbiasdness

Theorem 17. *For any $f_n \in \mathcal{B}_b(E_n)$, we have that:*

$$\mathbb{E}[\eta_n^N(f_n)] = \eta_n(f_n),$$
$$\mathbb{E}[\gamma_n^N(f_n)] = \gamma_n(f_n).$$

Proof. It suffices to prove that $\mathbb{E}[\gamma_n^N(f_n)] = \gamma_n(f_n)$. Let $\mathcal{F}_n^N$ be the filtration generated by particles at n. Then, by construction, we have that, for any $f_p \in \mathcal{B}_b(E_p)$:

$$\begin{aligned}\mathbb{E}\left[\eta_p^N(f_p) \mid \mathcal{F}_{p-1}^N\right] &= \frac{1}{N}\sum_{i=1}^N \mathbb{E}\left[f_p(\xi_p^{(i)}) \mid \mathcal{F}_{p-1}^N\right] = \mathbb{E}\left[f_p(\xi_p^1) \mid \mathcal{F}_{p-1}^N\right], \\ &= \frac{G_{p-1}(\xi_{p-1}^{(i)})}{\sum_{k=1}^N G_{p-1}(\xi_{p-1}^{(k)})} M_p(f_p(\xi_{p-1}^{(i)})) = \frac{\eta_{p-1}^N(Q_p(f_p))}{\eta_{p-1}^N(G_{p-1})}.\end{aligned}$$

Then, the tower property and the definition of γ_n^N give rise to:

$$\begin{aligned}\mathbb{E}[\gamma_n^N(f_n)] &= \mathbb{E}\left[\eta_n^N(f_n) \times \prod_{p=0}^{n-1} \eta_n^N(G_n) \mid \mathcal{F}_{n-1}^N\right] = \mathbb{E}\left[\frac{\eta_{n-1}^N(Q_n(f_n))}{\eta_{n-1}^N(G_{n-1})} \prod_{p=0}^{n-1} \eta_n^N(G_n)\right], \\ &= \mathbb{E}\left[\eta_{n-1}^N(Q_n(f_n)) \prod_{p=0}^{n-2} \eta_n^N(G_n)\right] = \mathbb{E}\left[\frac{\eta_{n-2}^N(Q_{n-1:n}(f_n))}{\eta_{n-2}^N(G_{n-2})} \prod_{p=0}^{n-2} \eta_n^N(G_n)\right],\end{aligned}$$

and so on. This gives rise to:

$$\begin{aligned}\mathbb{E}[\gamma_n^N(f_n)] &= \mathbb{E}\left[\eta_0^N(Q_{0:n}(f_n))\right] = \mathbb{E}\left[Q_{0:n}(f_n(\xi_0^{(1)}))\right], \\ &= \gamma_0 Q_{0:n}(f_n) = \gamma_n(f_n).\end{aligned}$$

□

3.6.2 L^2-bound

For any $\mu, \nu \in \mathcal{P}(E)$ and $f \in \mathcal{B}_b(E)$, we first define the following distance (Rebeschini and Van Handel, 2015):

$$d(\mu,\nu) := \sup_{\|f\|_\infty \leq 1} \sqrt{\mathbb{E}\left[(\mu(f)-\nu(f))^2\right]}, \tag{3.64}$$

where the supremum is taken over test functions satisfying $\|f\|_\infty := \sup_{x\in E} =\mid f(x) \mid \leq 1$. Recall that we can write interacting particle systems at every time step:

$$\eta_{n+1}^N = \Psi_{G_n}(\eta_n^N) M_{n+1},$$

with $\Psi_{G_n}(\eta_n^N(\xi_n)) := \sum_{i=1}^N \frac{G_n(\xi_n^{(i)})}{\sum_{j=1}^N G_n(\xi_n^{(j)})} \delta_{\xi_n^{(i)}}$ and an initial distribution $\hat{\eta}_0^N = \eta_0$.

Lemma 9. *For any Markov kernel M on $(E,\mathcal{E})$, any $\mu,\nu \in \mathcal{P}(E)$ and any $f \in \mathcal{B}_b(E)$, we have that:*

$$d(\mu M, \nu M) \leq d(\mu,\nu).$$

Proof. Notice that for $\|f\|_\infty \leq 1$:

$$\begin{aligned}\mid M(f) \mid &= \left| \int M(x',dx) f(x) \right| \leq \int M(x',dx) \mid f(x) \mid, \\ &\leq \int M(x',dx) \|f\|_\infty = \|f\|_\infty \leq 1,\end{aligned}$$

holds by Jensen's inequality, and this implies that $\|M(f)\|_\infty \leq 1$. Then we have that:

$$\begin{aligned}d(\mu M, \nu M) &= \sup_{\|f\|_\infty \leq 1} \sqrt{\mathbb{E}\left[(\mu M(f) - \nu M(f))^2\right]}, \\ &\leq \sup_{\|f\|_\infty \leq 1} \sqrt{\mathbb{E}\left[(\mu(f)-\nu(f))^2\right]}, \\ &= d(\mu,\nu).\end{aligned}$$

□

Lemma 10. *Assume that there exists $c \in (0,1)$ such that $c \leq G(x) \leq c^{-1}$ for all $x \in E$. Then for any*

$\mu, \nu \in \mathcal{P}(E)$ and any $f \in \mathcal{B}_b(E)$, we have that:

$$d\left(\Psi_G(\mu), \Psi_G(\nu)\right) \leq 2c^{-2} d\left(\mu, \nu\right).$$

Proof. For $f \in \mathcal{B}_b(E)$, we have that:

$$\begin{aligned}
\Psi_G(\mu)(f) - \Psi_G(\nu)(f) &= \frac{\mu\left(fG\right)}{\mu(G)} - \frac{\nu(fG)}{\nu(G)}, \\
&= \frac{\mu(fG) - \nu(fG)}{\mu(G)} + \frac{\nu(fG)}{\mu(G)} - \frac{\nu(fG)}{\nu(G)}, \\
&= \frac{\|G\|_\infty}{\mu(G)} \left[\mu\left(\frac{fG}{\|G\|_\infty}\right) - \nu\left(\frac{fG}{\|G\|_\infty}\right)\right] + \frac{\nu(fG)\,\|G\|_\infty}{\mu(G)\nu(G)} \left[\nu\left(\frac{G}{\|G\|_\infty}\right) - \mu\left(\frac{G}{\|G\|_\infty}\right)\right].
\end{aligned}$$

Notice that $\mu(G)^{-1} \leq c^{-1}$ and $\frac{\mu(fG)}{\mu(G)} \leq 1$ for any $\mu \in \mathcal{P}(E)$ since the potential function G is positive. Therefore, we have:

$$\Psi_G(\mu)(f) - \Psi_G(\nu)(f) \leq c^{-2}\left[\mu\left(cfG\right) - \nu\left(cfG\right)\right] + c^{-2}\left[\nu\left(cG\right) - \mu\left(cG\right)\right].$$

For $\|f\|_\infty \leq 1$, we have that $\|cfG\|_\infty \leq \|f\|_\infty \|G\|_\infty \leq \|G\|_\infty \leq 1$, and then triangle inequality implies:

$$\begin{aligned}
\mathbb{E}\left[\left|\Psi_G(\mu)(f) - \Psi_G(\nu)(f)\right|^2\right] &\leq \sup_{\|f\|_\infty \leq 1} \mathbb{E}\left[\left|c^{-2}\left[\mu\left(cfG\right) - \nu\left(cfG\right)\right]\right|^2\right] + \sup_{\|f\|_\infty \leq 1} \mathbb{E}\left[\left|c^{-2}\left[\nu\left(cG\right) - \mu\left(cG\right)\right]\right|^2\right], \\
&\leq 2c^{-4} \sup_{\|f\|_\infty \leq 1} \mathbb{E}\left[\left|\left[\mu\left(f\right) - \nu\left(f\right)\right]\right|^2\right].
\end{aligned}$$

The claim follows immediately. □

Theorem 18. *Assume that there exists $c \in (0,1)$ such that $c \leq G(x) \leq c^{-1}$ for all $x \in E$ and $\epsilon_n = 0$ for any n in Algorithm 13. Then for any $f \in \mathcal{B}_b(E_n)$, we have that:*

$$d\left(\eta_n^N, \eta_n\right) \leq \frac{C_n}{\sqrt{N}},$$

where the constant C_n does not depend on N but on n.

Proof. Assume that the statement holds at time $n-1$. Then, as we mentioned, we can write $\eta_n^N - \eta_n$ as:

$$\begin{aligned}
\eta_n^N - \eta_n &= \left[\eta_n^N - \Phi_n(\eta_{n-1}^N)\right] + \left[\Phi_n(\eta_{n-1}^N) - \Phi_n(\eta_{n-1})\right] \\
&= \left[\eta_n^N - \Phi_n(\eta_{n-1}^N)\right] + \left[\Psi_{G_{n-1}}(\eta_{n-1}^N)M_n - \Psi_{G_{n-1}}(\eta_{n-1})M_n\right].
\end{aligned}$$

From Lemma 9 and Lemma 10, we first obtain:

$$\begin{aligned}
\sup_{\|f\|_\infty \leq 1} \left\|\Psi_{G_{n-1}}(\eta_{n-1}^N)M_n(f_n) - \Psi_{G_{n-1}}(\eta_{n-1})M_n(f_n)\right\|_2 &\leq 2c^{-2} \sup_{\|f\|_\infty \leq 1} \left\|\eta_{n-1}^N M_n(f_n) - \eta_{n-1} M_n(f_n)\right\|_2, \\
&\leq 2c^{-2} \sup_{\|f\|_\infty \leq 1} \left\|\eta_{n-1}^N(f_n) - \eta_{n-1}(f_n)\right\|_2,
\end{aligned}$$

so that $d\left(\Psi_{G_{n-1}}(\eta_{n-1}^N)M_n, \Psi_{G_{n-1}}(\eta_{n-1})M_n\right) \leq 2c^{-2}d\left(\eta_{n-1}^N, \eta_{n-1}\right)$.

By construction, we have $\mathbb{E}\left[\eta_n^N(f) \mid \mathcal{F}_{n-1}^N\right] = \Phi_n(\eta_{n-1}^N)$ so that $\mathbb{E}\left[\sqrt{N}\left(\eta_n^N(f) - \Phi_n(\eta_{n-1}^N)(f)\right)\right] = \mathbb{E}\left[\mathbb{E}\left[\sqrt{N}\left(\eta_n^N(f) - \Phi_n(\eta_{n-1}^N)(f)\right) \mid \mathcal{F}_{n-1}^N\right]\right] = 0$ holds. Then, conditioned on $\mathcal{F}_{n-1}^N$, the Marcinkiewicz-Zygmund inequality yields that there exists a positive constant c' such that

$$d\left(\eta_n^N, \Phi_n(\eta_{n-1}^N)\right) \leq \frac{c'}{\sqrt{N}},$$

Collecting our estimates, we have:

$$d\left(\eta_n^N, \eta_n\right) \leq \frac{c'}{\sqrt{N}} + 2c^{-2}d\left(\eta_{n-1}^N, \eta_{n-1}\right).$$

At time $n = 0$, we have that:

$$\sup_{\|f\|_\infty \leq 1} \left\|\eta_0^N(f_0) - \eta_0(f_0)\right\|_2 \leq \frac{1}{\sqrt{N}},$$

see Appendix E. Thus the result follows by induction. □

3.6.3 Central limit theorem

We first analyse the local sampling errors associated with the interacting particle systems. Recall that using the semigroup $Q_{p:n}$ in (3.41), we can write $\gamma_n = \gamma_p Q_{p:n}$ for $p < n$. Therefore we can decompose $\gamma_n^N - \gamma_n$ as:

$$\gamma_n^N - \gamma_n = \sum_{p=1}^{n} \left[\gamma_p^N Q_{p:n} - \gamma_{p-1} Q_{p-1:n}\right] + \gamma_0^N Q_{0:n} - \gamma_n. \tag{3.65}$$

Also, for any $p \geq 1$, (3.43) gives rise to:

$$\begin{aligned}\gamma_{p-1}^N Q_p \eta_n(f_n) &= \gamma_{p-1}^N(1)\eta_{p-1}^N Q_p = \gamma_{p-1}^N(1)\eta_{p-1}^N(G_{p-1})\Phi_n(\eta_{n-1}^N),\\ &= \gamma_{p-1}^N(1)\frac{\gamma_{p-1}^N(G_{p-1})}{\gamma_{p-1}^N(1)}\Phi_n(\eta_{n-1}^N) = \gamma_{p-1}^N(G_{p-1})\Phi_n(\eta_{n-1}^N) = \gamma_p^N(1)\Phi_n(\eta_{n-1}^N).\end{aligned}$$

Using this we have that:

$$\begin{aligned}\gamma_p^N Q_{p:n} - \gamma_{p-1} Q_{p-1:n} &= \gamma_p^N Q_{p:n} - \gamma_{p-1} Q_p Q_{p:n} = \gamma_p^N Q_{p:n} - \gamma_p^N(1)\Phi_n(\eta_{n-1}^N)Q_{p:n},\\ &= \gamma_p^N(1)\left(\eta_p^N - \Phi_n(\eta_{n-1}^N)\right)Q_{p:n}.\end{aligned}$$

As a result, we obtain:

$$W_n^{\gamma,N} := \sqrt{N}\left(\gamma_n^N - \gamma_n\right) = \sum_{p=0}^{n} \gamma_p^N(1) W_p^N Q_{p:n}, \tag{3.66}$$

where we have defined $\left(W_p^N\right)_{p\geq 0}$ with $W_0^N := \sqrt{N}\left(\eta_p^N - \eta_0\right)$ as follows, for $p \geq 1$:

$$W_p^N := \sqrt{N}\left(\eta_p^N - \Phi_n(\eta_{n-1}^N)\right). \tag{3.67}$$

Notice that $\Phi_n(\eta_{n-1}^N) = K_{n,\eta_{n-1}^N}$ under the assumption $\epsilon_n = 0$ for any n. We then have the following.

Lemma 11. *(Del Moral, 2004, Corollary 9.3.1)*

Assume that $\epsilon_n = 0$ for any n in Algorithm 13. Then for any $f_n \in \mathcal{B}_b(E_n)$ and fixed $n \geq 0$, the process $\left\{W_n^N\right\}_{n\geq 0}$ converges in law as $N \to \infty$ to a sequence of independent centered Gaussian random fields $\{W_n\}_{n\geq 0}$ such that:

$$\mathbb{E}\left[W_n(f_n)^2\right] = \eta_n\left[f_n - \eta_n(f_n)^2\right].$$

Next we consider the process $\left\{W_n^{\eta,N}\right\}_{n\geq 0}$ such that for any $f_n \in \mathcal{B}_b(E_n)$, $W_n^{\eta,N}(f_n) := \sqrt{N}\left[\eta_n^N - \eta_n\right](f_n)$. Also notice that $\gamma_n\left(f_n - \eta_n(f_n)\right) = 0$ holds from the law of total expectation. Therefore, we have that:

$$\begin{aligned} W_n^{\eta,N}(f_n) &:= \sqrt{N}\left[\eta_n^N - \eta_n\right](f_n) = \sqrt{N}\frac{\gamma_n^N\left(f_n - \eta_n(f_n)\right)}{\gamma_n^N(1)}, \\ &= \sqrt{N}\frac{\left(\gamma_n^N - \gamma_n\right)\left(f_n - \eta_n(f_n)\right)}{\gamma_n^N(1)} = \frac{W_n^{\gamma,N}\left(f_n - \eta_n(f_n)\right)}{\gamma_n^N(1)}. \end{aligned} \tag{3.68}$$

Using these decompositions and Lemma 11 we can show the following.

Theorem 19. *Assume that $\epsilon_n = 0$ for any n in Algorithm 13. Then for any $f_n \in \mathcal{B}_b(E_n)$ and fixed $n \geq 0$, the process $\left\{W_n^{\gamma,N}\right\}_{n\geq 0}$ and $\left\{W_n^{\eta,N}\right\}_{n\geq 0}$ converge in law as $N \to \infty$ to a sequence of Gaussian random fields $\{W_n^{\gamma}\}_{n\geq 0}$ and $\{W_n^{\eta}\}_{n\geq 0}$ which are defined respectively as:*

$$\begin{aligned} W_n^{\gamma}(f_n) &:= \gamma_n(1)\sum_{p=0}^{n} W_p\left(P_{p:n}\left(f_n\right)\right), \\ W_n^{\eta}(f_n) &:= \sum_{p=0}^{n} W_p\left(P_{p:n}\left(f_n - \eta_n(f_n)\right)\right). \end{aligned}$$

Proof. From the strong law of large numbers, we have that $\gamma_n^N(1) \to \gamma_n(1)$ w.p.1 as $N \to \infty$. Then Lemma 11 and Slutsky's theorem imply $W_n^{\gamma}(f_n) \to \sum_{p=0}^{n}\gamma_p(1)W_p\left(Q_{p:n}(f_n)\right)$ in distribution as $N \to \infty$. Then the result follows from $Q_{p:n} = \frac{\gamma_n(1)}{\gamma_p(1)}P_{p:n}$ (3.46). The rest of the claim follows immediately. □

Theorem 20. *Assume that $\epsilon_n = 0$ for any n in Algorithm 13. Then for any $f_n \in \mathcal{B}_b(E_n)$ and fixed $n \geq 0$, we have that:*

$$\begin{aligned} \lim_{N\to\infty} N\mathbb{E}\left[\eta_n^N(f_n) - \eta_n(f_n)\right] &= -\sum_{p=0}^{n}\eta_p\left(P_{p:n}(1)P_{p:n}\left(f_n - \eta_n(f_n)\right)\right), \\ \lim_{N\to\infty} N\mathbb{V}\left[\eta_n^N(f_n) - \eta_n(f_n)\right] &= \sum_{p=0}^{n}\eta_p\left(P_{p:n}\left(f_n - \eta_n(f_n)\right)^2\right). \end{aligned}$$

Proof. Recall that $W_n^{\eta,N}(f_n)$ can be expressed as $W_n^{\eta,N}(f_n) := \sqrt{N}\left[\eta_n^N - \eta_n\right](f_n) = \frac{W_n^{\gamma,N}(f_n - \eta_n(f_n))}{\gamma_n^N(1)}$

so that:

$$N\left(\eta_n^N(f_n)-\eta_n(f_n)\right)=\sqrt{N}\frac{\gamma_n(1)}{\gamma_n^N(1)}W_n^{\gamma,N}\left(\frac{f_n-\eta_n(f_n)}{\gamma_n(1)}\right)$$
$$=\sqrt{N}\left[\frac{\gamma_n(1)}{\gamma_n^N(1)}-1\right]W_n^{\gamma,N}\left(\frac{f_n-\eta_n(f_n)}{\gamma_n(1)}\right)+\sqrt{N}W_n^{\gamma,N}\left(\frac{f_n-\eta_n(f_n)}{\gamma_n(1)}\right).$$

From Theorem 17, $\mathbb{E}\left[W_n^{\eta,N}\right]=0$ holds, and this implies $\mathbb{E}\left[\sqrt{N}W_n^{\eta,N}\left(\frac{f_n-\eta_n(f_n)}{\gamma_n(1)}\right)\right]=0$. Notice that $W_n^{\gamma,N}:=\sqrt{N}\left(\gamma_n^N-\gamma_n\right)$ and:

$$\left[\frac{\gamma_n(1)}{\gamma_n^N(1)}-1\right]=-\frac{1}{\gamma_n^N(1)}\left[\gamma_n^N(1)-\gamma_n(1)\right]=-\frac{1}{\sqrt{N}}\frac{W_n^{\gamma,N}(1)}{\gamma_n^N(1)}.$$

As a result, we obtain:

$$N\mathbb{E}\left[\eta_n^N(f_n)-\eta_n(f_n)\right]=-\mathbb{E}\left[\frac{W_n^{\gamma,N}(1)}{\gamma_n^N(1)}W_n^{\gamma,N}\left(\frac{f_n-\eta_n(f_n)}{\gamma_n(1)}\right)\right],$$
$$=-\frac{1}{\gamma_n(1)}\mathbb{E}\left[W_n^{\gamma,N}(1)W_n^{\eta,N}\left(f_n\right)\right].$$

here we again used Theorem 17. From Slutsky's theorem, dominated convergence theorem and Theorem 19, we can show that:

$$\lim_{N\to\infty}N\mathbb{E}\left[\eta_n^N(f_n)-\eta_n(f_n)\right]=-\sum_{p=0}^{n}\mathbb{E}\left[W_p\left(P_{p:n}\left(1\right)\right)W_p\left(P_{p:n}\left(f_n-\eta_n(f_n)\right)\right)\right],$$
$$=-\sum_{p=0}^{n}\eta_p\left(P_{p:n}\left(1\right)P_{p:n}\left(f_n-\eta_n(f_n)\right)\right),$$

here we used Lemma 11. From basic decomposition of the variance, we have:

$$\mathbb{V}\left[\eta_n^N(f_n)-\eta_n(f_n)\right]=\mathbb{E}\left[\left(\eta_n^N(f_n)-\eta_n(f_n)\right)^2\right]-\left(\mathbb{E}\left[\eta_n^N(f_n)-\eta_n(f_n)\right]\right)^2.$$

We already know that $\left(\mathbb{E}\left[\eta_n^N(f_n)-\eta_n(f_n)\right]\right)=\mathcal{O}\left(N^{-2}\right)$ thus can be ignored. We also obtain $\mathbb{E}\left[\left(\eta_n^N(f_n)-\eta_n(f_n)\right)^2\right]=\frac{1}{N}\mathbb{E}\left[\left(W_n^{\eta,N}(f_n)\right)^2\right]$. Therefore, again from Slutsky's theorem, dominated convergence theorem and Theorem 19, we have:

$$\lim_{N\to\infty}N\mathbb{V}\left[\eta_n^N(f_n)-\eta_n(f_n)\right]=\sum_{p=0}^{n}\mathbb{E}\left[\sum_{p=0}^{n}W_p\left(P_{p:n}\left(f_n-\eta_n(f_n)\right)\right)^2\right],$$
$$=\sum_{p=0}^{n}\eta_p\left(P_{p:n}\left(f_n-\eta_n(f_n)\right)\right)^2,$$

here again we used Lemma 11. □

4 Hidden Markov Models and Particle Filters

4.1 Introduction

Owing to their rich structure, hidden Markov models (HMMs) are being routinely used in such diverse disciplines as finance (Mamon and Elliott, 2007), speech recognition (Gales and Young, 2008), epidemiology (Green and Richardson, 2002), biology (Yoon, 2009), and signal processing (Crouse et al., 1998). This section provides some basic and detailed results of HMMs. Since HMMs involve intractable densities with unknown parameters in general, approximation methods and statistical inference based on SMC for HMMs are documented.

4.2 Basics of Hidden Markov Models

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space on which we have defined two stochastic processes $\{x_n; n \in \mathbb{N}\}$ and $\{y_n; n \in \mathbb{N}\}$. The process $\{x_n\}$ is a latent Markov process which takes values in $\mathbb{X}$. Also, let $\mathcal{B}(\mathbb{X})$ be the Borel σ-algebra on $\mathbb{X}$. Then the probability transition kernel $f_\theta : \mathbb{X} \times \mathcal{B}(\mathbb{X}) \to [0, 1]$ of $\{x_n\}$ is such that:

$$\mathbb{P}(x_n \in A \mid x_{n-1}) = \int_A f_\theta(dx_n \mid x_{n-1}),\ A \in \mathcal{B}(\mathbb{X}), \tag{4.1}$$

Similarly, let $\{y_n\}$ be an observation process which is conditionally independent of $x_n = x$ over $n \geq 0$ and have the marginal distribution $g_\theta : \mathbb{X} \times \mathcal{B}(\mathbb{Y}) \to [0, 1]$ such that:

$$\mathbb{P}(y_n \in B \mid x_n) = \int_B g_\theta(dy_n \mid x_n),\ B \in \mathcal{B}(\mathbb{Y}). \tag{4.2}$$

Then *Hidden Markov Models* (HMMs) (also known as *State Space Models*) are defined as the bivariate stochastic process $(x_n, y_n)_{n\in\mathbb{N}}$. That is, HMMs are $(\mathbb{X} \times \mathbb{Y}, \mathcal{B}(\mathbb{X}) \otimes \mathcal{B}(\mathbb{Y}))$-measurable Markov chains. We adopt a parametric setting with $\theta \in \Theta \subseteq \mathbb{R}^d$, for some $d \in \mathbb{N}$. Assume that $f_\theta(dx_n \mid x_{n-1})$ and $g_\theta(dy_n \mid x_n)$ admit densities w.r.t. the dominating measures denoted as dx and dy, with abuse of notation. That is:

$$\mathbb{P}(x_n \in dx_n \mid x_{n-1}) = f_\theta(x_n \mid x_{n-1})dx_n, \tag{4.3}$$

$$\mathbb{P}(y_n \in dy_n \mid x_{n-1}) = g_\theta(y_n \mid x_n)dy_n. \tag{4.4}$$

In the context of HMMs, such models are often called *fully dominated models* (Cappé et al., 2005; Douc et al., 2014). Then for any $A \in \mathcal{B}(\mathbb{X})$, one can define the joint Markov density k_θ w.r.t. the product measure $(dx \otimes dy)$ such as:

$$k_\theta(x_n, y_n \mid x_{n-1}) := f_\theta(x_n \mid x_{n-1})g_\theta(y_n \mid x_n), \tag{4.5}$$

for $x_n, x_{n+1} \in \mathbb{X} \times \mathbb{X}$ and $y_n \in \mathbb{Y}$. Following closely Doucet and Johansen (2009), we provide several examples of HMMs to facilitate our study.

Example 14. *Stochastic Volatility model.*

Stochastic Volatility (SV) models, first studied in Taylor (1982), might be the most used class among HMMs. See Shephard and Andersen (2009) and Durbin and Koopman (2012, section 9) for details of SV models and their extension.

Let $\{y_n\}$ denote the first difference of a particular series of asset log prices. Such prices might be of stocks, bonds, foreign currencies. A common specification of $\{y_n\}$ will be given by:

$$y_n = \beta \exp(x_n/2) w_n,$$

where $w_n \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$. The latent process $\{x_n\}$ can be thought of as the unobserved log $-$volatility. A standard assumption on $\{x_n\}$ will be that $\{x_n\}$ follows a first order autoregression:

$$x_n = \alpha x_{n-1} + v_n,$$

where $\mid \alpha \mid < 1$ and $v_n \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_x^2)$. First, we note that such SV models can be understood as a natural discrete time analogue of the continuous models, such as models studied by Hull and White (1987). Also, it should be emphasised that although SV models are Gaussian, they are not linear models. Therefore, one has to resort to some sophisticated methods to approximate important quantities such as likelihood and posterior of SV models. We give a simulation of the SV model with $n = 1000$ and $(\alpha, \sigma_x, \beta) = (0.9, \sqrt{0.1}, 0.8)$ in the following Figure 2.

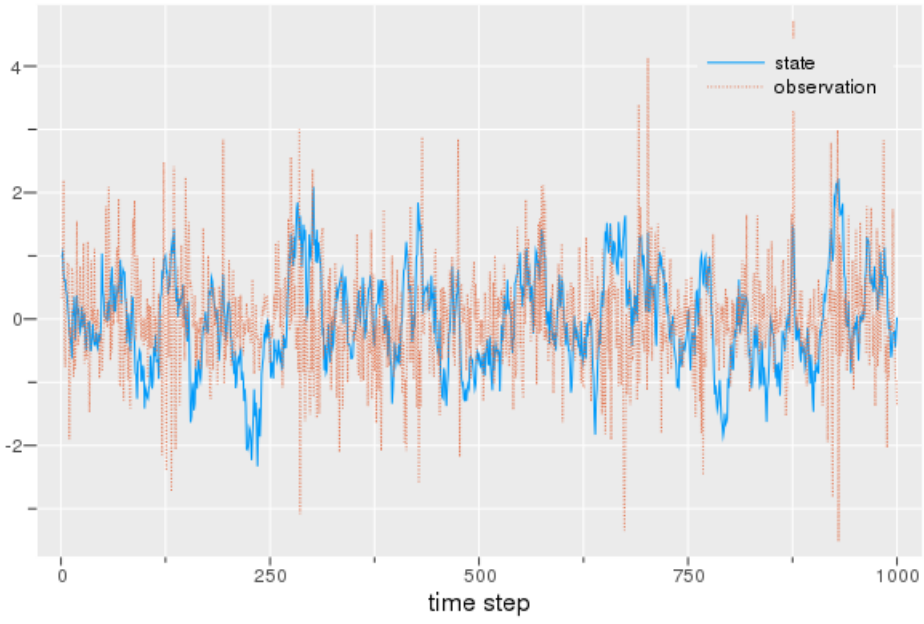


Figure 2: A simulation of the stochastic volatility model described in Example 14 with parameters $(\alpha, \sigma_x, \beta) = (0.9, \sqrt{0.1}, 0.8)$.

Example 15. *Partially observed SDEs with error.*

We consider a d-dimensional diffusion model. Suppose that stochastic process $X = \{X_t; 0 \leq t \leq T\}$

is obtained by the solution to the following time-homogeneous stochastic differential equation (SDE):

$$dX_t = b_\theta(X_t)dt + \sigma_\theta(X_t)dW_t,\ X_0 = x_0 \in \mathbb{R}^p,\ t \in [0, T]$$

driven by the Brownian motion $\{W_t; 0 \leq t \leq T\}$, where $b : \mathbb{R}^d \mapsto \mathbb{R}^d$ and $\sigma : \mathbb{R}^d \mapsto \mathbb{R}^{d\times d}$, that might depend on also some parameter $\theta \in \Theta \subseteq \mathbb{R}^p$ to be estimated. We make the assumptions (linear growth and Lipschitz continuity) to guarantee uniqueness of a weak solution, see, for instance Øksendal (2003).

We assume that the process can be observed at only discrete time instances $1 \leq m \leq n$, $0 \leq t_1 < t_2 < \cdots t_n$ with error $y_n \mid x_n = g_\theta(dy_n \mid x_n)$.This class of models have been particularly used in finance and financial econometrics to capture market microstructure noise. See Aït-Sahalia et al. (2005); Ait-Sahalia and Yu (2008); Hansen and Lunde (2006) for instance. For convenience, we also write $x_m = X_{t_m}, 0 \leq m \leq n$. Note the setting above can be still seen as a special case of HHMs, but time is continuous. Again, to facilitate our study, we heuristically write the distribution of x_n given x_{n-1} as $f_\theta(dx_n \mid x_{n-1})$ and assume that it has the density w.r.t. some dominating measure denoted generically as dx. Also we assume that $g_\theta(dy \mid x)$ admit densities w.r.t. some dominating measure denoted dy.

Note that, as for SDEs case, $f_\theta(x_n \mid x_{n-1})$ *is itself intractable* in general, and this yields some unique difficulties compared with basic HMMs. Then, common choice of $g_\theta(y_n \mid x_n)$ will be Gaussian, that is $y_n = x_n + \epsilon_n$ where $\epsilon_n \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$. We give a simulation of the such $1-$dimensional SDEs with error model with the specification to illustrate features of such models :

$$\begin{aligned} dX_t &= 0.5(0.7 - X_t)dt + 0.5dW_t, \\ y_n &= x_n + \epsilon_n,\ \epsilon_n \overset{i.i.d.}{\sim} \mathcal{N}(0, 1), \end{aligned}$$

in the following Figure 3.

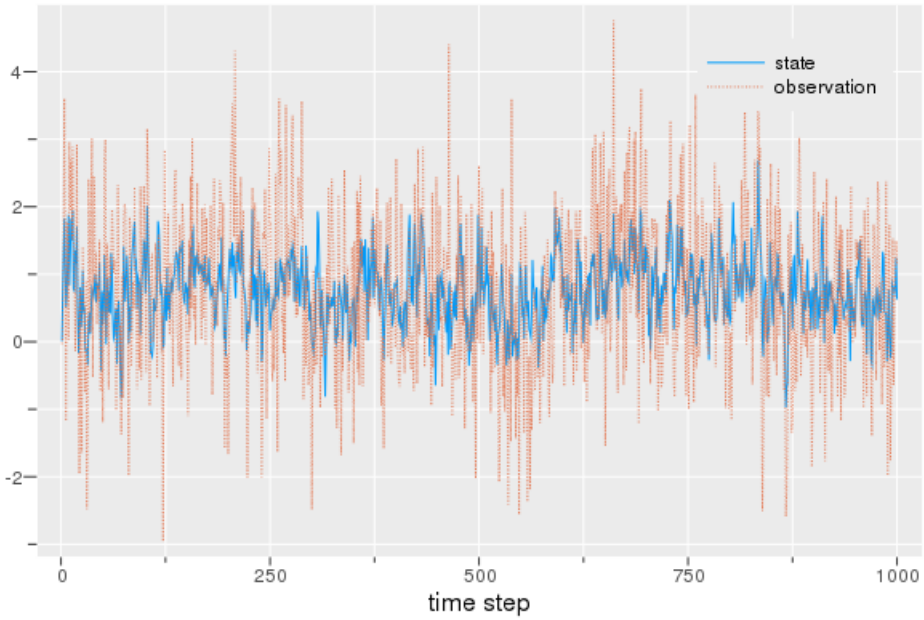


Figure 3: A simulation of Partially observed 1−dimensional SDEs with error model described in Example 15 with $b_\theta = 0.5(0.7 - X_t), \sigma_\theta = 0.5$ and $y_n = x_n + \epsilon_n,\ \epsilon_n \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$.

Henceforth we make use of the notation $a_{i:j} := (a_i, \cdots, a_j)$, for integers $i \leq j$, for a given sequence $\{a_m\}$. Then estimation problems for HHMs involve the posterior distribution of $\{x_n\}$ given $y_{0:k}$. That is, we are interested in computing the conditional distributions:

$$\mathbb{P}(x_k \in dx_k \mid y_{0:n}). \tag{4.6}$$

When $n = k$, (4.6) becomes the *filtering* distribution, when $0 \leq k < n$ it corresponds to *smoothing* distribution, when $k > n$ it is called *prediction* distribution. Note that the following development will be done under the assumption that the parameter θ is known. Therefore, we will drop θ from the expressions, that is, we will write $f(x_n \mid x_{n-1})$ instead of $f_\theta(x_n \mid x_{n-1})$ for instance.

Definition 30. (Joint smoothing, Filtering, Smoothing, Prediction densities)

i) *Joint smoothing density:* $p(x_{0:n} \mid y_{0:n})$ *for* $n \geq 0$.

ii) *Smoothing density:* $p(x_k \mid y_{0:n})$ *for* $0 \leq k \leq n$.

iii) *Filtering density:* $p(x_n \mid y_{0:n})$ *for* $n \geq 0$.

iv) *Prediction density:* $p(x_{n+1} \mid y_{0:n})$ *for* $n \geq 0$.

Due to Markovian structure of HMMs and the Bayes' theorem, one can easily derive recursions for such posterior densities. We first begin with the recursion for the joint smoothing density (JSD). Since we have considered fully dominated HMMs, Markov property gives rise to:

$$p(x_{0:n}, y_{0:n}) = p(x_{0:n-1}, y_{0:n-1}) f(x_n \mid x_{n-1}) g(y_n \mid x_n), \tag{4.7}$$

where $p(x_{0:n}, y_{0:n})$ denotes the joint density of $(x_{0:n}, y_{0:n})$ for all $n \geq 1$. Then, from the Bayes' theorem, we obtain:

$$p(x_{0:n} \mid y_{0:n}) = p(x_{0:n-1} \mid y_{0:n-1}) \frac{f(x_n \mid x_{n-1}) g(y_n \mid x_n)}{p(y_n \mid y_{0:n-1})}, \tag{4.8}$$

$$p(y_n \mid y_{0:n-1}) = \int p(x_{0:n-1} \mid y_{0:n-1}) f(x_n \mid x_{n-1}) g(y_n \mid x_n) dx_{n-1:n}. \tag{4.9}$$

As for the filtering density $p(x_n \mid y_{0:n})$, one can also derive the recursion in the same manner:

$$\begin{aligned} p(x_n \mid y_{0:n}) &= \frac{p(y_n \mid y_{0:n-1}, x_n) p(x_n \mid y_{0:n-1})}{p(y_n \mid y_{0:n-1})}, \\ &= \frac{g(y_n \mid x_n) p(x_n \mid y_{0:n-1})}{p(y_n \mid y_{0:n-1})}, \end{aligned} \tag{4.10}$$

where $p(y_n \mid y_{0:n-1}) = \int g(y_n \mid x_n) p(x_n \mid y_{0:n-1}) dx_n$. Again, we have used Markovian structure of HMMs and the Bayes' theorem. Notice that the prediction density appears in the filtering recursion (4.10), and the prediction recursion can be obtained as:

$$\begin{aligned} p(x_n \mid y_{0:n-1}) &= \int p(x_n, x_{n-1} \mid y_{0:n-1}) dx_{n-1}, \\ &= \int f(x_n \mid x_{n-1}) p(x_{n-1} \mid y_{0:n-1}) dx_{n-1}. \end{aligned} \tag{4.11}$$

Again notice that the filtering appears in the prediction recursion (4.11). Using (4.11), alternatively, (4.10) becomes:

$$p(x_n \mid y_{0:n}) = \frac{g(y_n \mid x_n) \int f(x_n \mid x_{n-1}) p(x_{n-1} \mid y_{0:n-1}) dx_{n-1}}{\int g(y_n \mid x_n) f(x_n \mid x_{n-1}) p(x_{n-1} \mid y_{0:n-1}) dx_{n-1:n}}. \tag{4.12}$$

Again notice that In the context of the optimal filtering problem, the recursion (4.11) is known as the *prediction step* and (4.12) is known as *filtering step*. Then it is clear to see that one can calculate the filtering density $p(x_n \mid y_{0:n})$ for any n by iterating the prediction step and the filtering step.

One of the interesting facts about HMMs is that, given $y_{0:n}$, the process $\{x_k\}_{k=0}^n$ is inhomogeneous Markov process. Under weak assumptions on a space $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$ (it has to be Polish space), the same holds true for the time-reversed chains. That is, for $A \in \mathcal{B}(\mathbb{X})$, define the backward kernel:

$$B(x_t \in A \mid x_{t+1}) := \mathbb{P}\left(x_t \in A \mid x_{t+1}, y_{0:n}\right), \tag{4.13}$$

then (4.13) is the time-reversed Markov kernel (Cappé et al., 2005, Proposition 3.3.6). Critically, since we have assumed that HMMs being considered are fully dominated, one can obtain an explicit

expression for the backward kernel:

$$\begin{aligned} p(x_k \mid x_{k+1}, y_{0:n}) &= p(x_k \mid x_{k+1}, y_{0:k}), \\ &= \frac{p(x_{k+1} \mid x_k, y_{0:k}) p(x_k \mid y_{0:k})}{\int p(x_{k+1} \mid x_k, y_{0:k}) p(x_k \mid y_{0:k}) dx_k}, \\ &= \frac{f(x_{k+1} \mid x_k) p(x_k \mid y_{0:k})}{\int f(x_{k+1} \mid x_k) p(x_k \mid y_{0:k}) dx_k}, \end{aligned} \tag{4.14}$$

where we used the conditional independence such that $p(x_k \mid x_{k+1}, y_{0:n}) = p(x_k \mid x_{k+1}, y_{0:k})$ holds for $n \geq k$. Also, notice that, for $0 \leq k \leq n$,

$$\begin{aligned} p(x_k \mid y_{0:n}) &= \int p(x_k \mid x_{k+1}, y_{0:n}) p(x_{k+1} \mid y_{0:n}) dx_{k+1}, \\ &= \int p(x_k \mid x_{k+1}, y_{0:k})) p(x_{k+1} \mid y_{0:n}) dx_{k+1}. \end{aligned} \tag{4.15}$$

Then combining (4.14) and (4.15) yields:

$$p(x_k \mid y_{0:n}) = p(x_k \mid y_{0:k}) \int \frac{f(x_{k+1} \mid x_k)}{p(x_{k+1} \mid y_{0:k})} p(x_{k+1} \mid y_{0:n}) dx_{k+1}, \tag{4.16}$$

$$p(x_{k+1} \mid y_{0:k}) = \int f(x_{k+1} \mid x_k) p(x_k \mid y_{0:k}) dx_k. \tag{4.17}$$

(4.16) implies that, to compute $\{p(x_k \mid y_{0:n})\}$, first one has to do the filtering forward and then compute $p(x_k \mid y_{0:n})$ backward. Thus, (4.16) is often called *Forward-Backward* recursion.

Although we have derived the recursions to obtain the posterior densities, except some simple models (such linear Gaussian models), it is not possible to compute these densities in closed-form. This difficulty motivates us to resort to some approximation methods. Sequential Monte Carlo (SMC) methods have been considered state of the art for tackling this kind of problems.

4.3 Hidden Markov Models and the Feynman-Kac Models

Recall that given the pair of a potential function and a Markov kernel (G_n, M_n), the Feynman-Kac prediction path models on the path space $E^n := \prod_{p=0}^n E_p$ equipped with the product $\mathcal{E}^n := \prod_{p=0}^n \mathcal{E}_p$ is given by $\mathbb{Q}(dx_{0:n}) := Z_n^{-1} \prod_{p=0}^{n-1} G_p(x_p) \mathbb{P}(dx_{0:n})$ where $\mathbb{P}(dx_{0:n}) := \mu_0(dx_0) \prod_{p=1}^n M_p(x_{p-1}, dx_p)$ and $Z_n := \mathbb{E}_{\mathbb{P}} \left[\prod_{p=0}^{n-1} G_p(x_p) \right]$.

To see the connection between the Feynman-Kac prediction path models and HMMs, take, for all $n \geq 0$, $E_n = \mathbb{X}$ and $\mathcal{E}_n = \mathcal{B}(\mathbb{X})$. For $A \in \mathcal{B}(\mathbb{X})$, set $M_n(x_{n-1}, A) = \int_A f(x_n \mid x_{n-1}) dx_n$ and $\mathbb{P}(Y_n \in B \mid X_n) = \int_B g(y_n \mid x_n) dy_n$ for $B \in \mathcal{B}(\mathbb{Y})$. For $h_n \in \mathcal{B}_b(\mathbb{X}^n)$, the prediction path model

becomes:

$$\begin{aligned}\mathbb{Q}(h_n) &= Z_n^{-1} \int h_n(x_{0:n}) \prod_{p=0}^{n-1} g(y_p \mid x_p) \prod_{k=0}^{n} f(x_k \mid x_{k-1}) dx_{0:k}, \\ &= p(y_{0:n-1}) \int h_n(x_{0:n}) p(x_{0:n}, y_{0:n-1}) dx_{0:n}, \\ &= \int h_n(x_{0:n}) p(x_{0:n} \mid y_{0:n-1}) dx_{0:n}.\end{aligned}$$

Thus, in the context of HMMs, $\mathbb{Q}(dx_{0:n})$ *is the (joint) predictive distribution* $p(dx_{0:n} \mid y_{0:n-1})$. In the same manner, it can be easily seen that, for $h_n \in \mathcal{B}_b(\mathbb{X}^n)$, $\hat{\mathbb{Q}}(h_n) = \int h_n(x_{0:n}) p(x_{0:n} \mid y_{0:n}) dx_{0:n}$ where $\hat{\mathbb{Q}}(dx_{0:n})$ is the Feynman-Kac updated path model. Thus $\hat{\mathbb{Q}}(dx_{0:n})$ *corresponds to the joint smoothing distribution* $p(dx_{0:n} \mid y_{0:n})$.

Next consider the flow of the time marginals of the Feynman-Kac models. That is, for $h_n \in \mathcal{B}_b(E_n)$, consider the following sequence of positive signed measures $\gamma_n(h_n) := \mathbb{E}_{\mathbb{P}}\left[h_n(x_n) \prod_{p=0}^{n-1} G_p(x_p)\right]$ and $\hat{\gamma}_n(h_n) := \mathbb{E}_{\mathbb{P}}\left[h_n(x_n) \prod_{p=0}^{n} G_p(x_p)\right]$. In the case of HMMs, again, take, for all $n \geq 0$, $E_n = \mathbb{X}$ and $\mathcal{E}_n = \mathcal{B}(\mathbb{X})$. For $A \in \mathcal{B}(\mathbb{X})$, set $M_n(x_{n-1}, A) = \int_A f(x_n \mid x_{n-1}) dx_n$ and $\mathbb{P}(Y_n \in B \mid X_n) = \int_B g(y_n \mid x_n) dy_n$ for $B \in \mathcal{B}(\mathbb{Y})$. For $h_n \in \mathcal{B}_b(\mathbb{X})$, $\gamma_n(h_n)$ becomes:

$$\begin{aligned}\gamma_n(h_n) &= \int h_n(x_n) \prod_{p=0}^{n-1} g(y_p \mid x_p) \prod_{k=0}^{n} f(x_k \mid x_{k-1}) dx_{0:k}, \\ &= \int h_n(x_n) p(x_{0:n}, y_{0:n-1}) dx_{0:n}, \\ &= p(y_{0:n-1}) \int h_n(x_n) p(x_{0:n} \mid y_{0:n-1}) dx_{0:n}, \\ &= \int h_n(x_n) p(x_n, y_{0:n-1}) dx_n.\end{aligned}$$

Thus, in the context of HMMs, $\gamma_n(h_n)$ *is related to the joint distribution of* $(x_n, y_{0:n-1})$. In the same manner, one can easily find that $\hat{\gamma}_n(h_n)$ *is related to the joint distribution of* $(x_n, y_{0:n})$. Therefore, the normalised version of time marginals of the Feynman-Kac models $\eta_n(h_n) := \gamma_n(h_n)/\gamma_n(1)$ and $\hat{\eta}_n(h_n) := \hat{\gamma}_n(h_n)/\hat{\gamma}_n(1)$ for $h_n \in \mathcal{B}_b(E_n)$ correspond to *the predictive distribution* $p(x_n \mid y_{0:n-1})$ *and the filtering distribution* $p(x_n \mid y_{0:n})$ with $E_n = \mathbb{X}$. Also it is clear to see that $\gamma_n(1) = p(y_{0:n-1})$ *and* $\hat{\gamma}_n(1) = p(y_{0:n})$.

In addition, recall that *the Boltzmann-Gibbs transformation* is the mapping $\Psi_{G_n} : \mathcal{P}(E_n) \to \mathcal{P}(E_n)$, defined for all $\mu_n \in \mathcal{P}(E_n)$ by:

$$\Psi_{G_n}(\mu_n)(dx_n) := \frac{G_n(x_n)\mu_n(dx_n)}{\mu_n(G_n)}.$$

To see the meaning of the Boltzmann-Gibbs transformation in the context of HMMs setting, recall

that $\eta_n(h_n) = \int h_n(x_n)p(x_n \mid y_{0:n-1})dx_n$. Then the Boltzmann-Gibbs transformation becomes:

$$\begin{aligned}\Psi_{G_n}(\eta_n)(dx_n) &= \frac{g(y_n \mid x_n)p(x_n \mid y_{0:n-1})dx_n}{\int g(y_n \mid x_n)p(x_n \mid y_{0:n-1})dx_n}, \\ &= p(dx_n \mid y_{0:n}).\end{aligned}$$

So, it is clear to see that *the Boltzmann-Gibbs transformation = Bayes theorem,* that is we have that $\Psi_{g_n}(\eta_n)(h_n) = \int h_n(x_n)p(x_n \mid y_{0:n})dx_n = \hat{\eta}_n(h_n)$. Using the transformation, we have that for $h_n \in \mathcal{B}_b(\mathbb{X})$:

$$\begin{aligned}\int h_n(x_n)p(x_n \mid y_{0:n}) &= \frac{\int h_n(x_n)g(y_n \mid x_n)p(x_n \mid y_{0:n-1})dx_n}{\int g(y_n \mid x_n)p(x_n \mid y_{0:n-1})dx_n} \\ &= \frac{\int h_n(x_n)g(y_n \mid x_n)\int f(x_n \mid x_{n-1})p(x_{n-1} \mid y_{0:n-1})dx_{n-1}dx_n}{\int g(y_n \mid x_n)\int f(x_n \mid x_{n-1})p(x_{n-1} \mid y_{0:n-1})dx_{n-1}dx_n}, \\ &= \frac{\int h_n(x_n)g(y_n \mid x_n)\hat{\eta}_{n-1}f(dx_n)}{\int g(y_n \mid x_n)\hat{\eta}_{n-1}f(dx_n)} = \Psi_{g_n}\left(\hat{\eta}_{n-1}f(h_n)\right).\end{aligned}$$

In the same manner, we have that $\eta_n(h_n) = \int h_n(x_n)p(x_n \mid y_{0:n-1})dx_n$. Then, observe that:

$$\begin{aligned}\int h_n(x_n)p(x_n \mid y_{0:n-1})dx_n &= \int h_n(x_n)\int f(x_n \mid x_{n-1})p(x_{n-1} \mid y_{0:n-1})dx_{n-1}dx_n, \\ &= \frac{\int h_n(x_n)\int f(x_n \mid x_{n-1})g(y_{n-1} \mid x_{n-1})p(x_{n-1} \mid y_{0:n-2})dx_{n-1}dx_n}{\int g(y_{n-1} \mid x_{n-1})p(x_{n-1} \mid y_{0:n-2})dx_{n-1}}, \\ &= \Psi_{g_{n-1}}(\eta_{n-1})f_n(h_n).\end{aligned}$$

Therefore, in the context of HMMs, the flow of the time marginals of the Feynman-Kac model can be considered as:

$$\eta_n \overset{Filtering/Bayes'\ rule}{\longrightarrow} \hat{\eta}_n = \Psi_{g_n}(\eta_n) \overset{prediction}{\longrightarrow} \eta_{n+1} = \hat{\eta}_n f_{n+1}.$$

We end this section by studying change of measures for HMMs. Again, take, for all $n \geq 0$, $E_n = \mathbb{X}$ and $\mathcal{E}_n = \mathcal{B}(\mathbb{X})$. For $A \in \mathcal{B}(\mathbb{X})$, set $M_n(x_{n-1}, A) = \int_A f(dx_n \mid x_{n-1})$ and $\mathbb{P}(Y_n \in B \mid X_n) = \int_B g(y_n \mid x_n)dy_n$ for $B \in \mathcal{B}(\mathbb{Y})$. Define the law on $(\mathbb{X}^n, \mathcal{B}(\mathbb{X}^n))$ such that:

$$\mathbb{H}(dx_{0:n}) = \mu(dx_0)\prod_{p=1}^{n} f(dx_p \mid x_{p-1}),$$

and set $G_n = g(y_n \mid x_n)$ for any $n \geq 0$. Then we have the Feynman-Kac path measure:

$$\begin{aligned}\mathbb{Q}(dx_{0:n}) &= p(y_{0:n-1})^{-1}\prod_{p=0}^{n-1} g(y_p \mid x_p)\mathbb{H}(dx_{0:n}), \\ p(y_{0:n-1}) &= \mathbb{E}_{\mathbb{H}}\left[\prod_{p=0}^{n-1} g(y_p \mid x_p)\right].\end{aligned}$$

Consider another collection of Markov chains $q(dx_n \mid x_{n-1})$ such that for any $x_{n-1} \in \mathbb{X}$, $f(dx_n \mid x_{n-1}) \ll q(dx_n \mid x_{n-1})$. Given this, as observed, we have:

$$\bar{\mathbb{P}}(dx_{0:n}) = \bar{\mu}_0(dx_0) \prod_{p=1}^{n} q(dx_p \mid x_{p-1}),$$
$$\bar{G}_p(x_{p-1}, x_p) = g(y_p \mid x_p) \times \frac{df(\cdot \mid x_{p-1})}{dq(\cdot \mid x_{p-1})}(x_p),$$

where $\mathbb{H} \ll \bar{\mathbb{P}}$. Then, as we will see later, *most SMC algorithms for HMMs can be considered as a special case of this framework*. That is, given $\mathbb{P}$ and $g_n(\cdot \mid \cdot)$ (and thus $\mathbb{Q}$), one can select arbitrary $\bar{\mathbb{P}}$ as long as $\mathbb{H} \ll \bar{\mathbb{P}}$ holds.

4.4 Particle Filter

In this section, we study how to approximate the filtering density $p(x_n \mid y_{0:n})$ or the joint smoothing density $p(x_{0:n} \mid y_{0:n})$ of HMMs via SMC. Applying SMC to the filtering problem of HMMs is commonly called the *particle filter*. As we studied, HMMs can be understood as a special case of Feynman-Kac models. To be precise, the potential function G_n is $g(y_n \mid x_n)$ and the Markov kernel M_n is $f(x_n \mid x_{n-1})$ in the context of HMMs. Therefore, applying particle approximation to HMMs immediately gives rise to an online procedure to approximate $p(x_n \mid y_{0:n})$ and $p(x_{0:n} \mid y_{0:n})$ for any n.

4.4.1 Bootstrap Filter

First, assume that our target is the joint smoothing density $p(x_{0:n} \mid y_{0:n})$. In the context of HMMs, the *Bootstrap Filter* presented in Gordon et al. (1993) has been intensively used. In this case, one just needs to set $\gamma_n(x_{0:n}) = p(x_{0:n}, y_{0:n})$ so that we also have $Z_n := \int \gamma_n(x_{0:n}) dx_{0:n} = \int p(x_{0:n}, y_{0:n}) dx_{0:n} = p(y_{0:n})$, and $\pi_n(x_{0:n}) := \frac{\gamma_n(x_{0:n})}{Z_n} = p(x_{0:n} \mid y_{0:n})$. In the bootstrap filter, the importance density $q(x_n \mid x_{0:n-1})$ is set as $f(x_n \mid x_{n-1})$. Given unweighted particle system $\left\{\tilde{x}^{(i)}_{0:n-1}, \frac{1}{N}\right\}_{i=1}^{N}$ for $p(x_{0:n-1} \mid y_{0:n-1})$, new particles $\{x_n^{(i)}\}_{i=1}^N$ are sampled from $f(x_n \mid \tilde{x}^{(i)}_{n-1})$ and set $x^{(i)}_{0:n} = \left(x_n^{(i)}, \tilde{x}^{(i)}_{0:n-1}\right)$. Notice that:

$$\begin{aligned} a_n(x_{0:n}) &:= \frac{\gamma_n(x_{0:n})}{\gamma_{n-1}(x_{0:n-1}) q(x_n \mid x_{0:n-1})} = \frac{p(x_{0:n}, y_{0:n})}{p(x_{0:n-1}, y_{0:n-1}) f(x_n \mid x_{n-1})}, \\ &= g(y_n \mid x_n). \end{aligned} \tag{4.18}$$

Also, $x^{(i)}_{0:n} = \left(x_n^{(i)}, \tilde{x}^{(i)}_{0:n-1}\right)$ is approximately distributed according to $p(x_{0:n-1} \mid y_{0:n-1}) f(x_n \mid x_{n-1})$ due to the resampling at time $n-1$. Therefore, the unnormalised weight is now given by:

$$w_n(x_{0:n}) = a_n(x_{0:n}) = g(y_n \mid x_n). \tag{4.19}$$

Next we calculate normalised weights as:

$$W_n^{(i)} := \frac{w_n(x^{(i)}_{0:n})}{\sum_{j=1}^N w_n(x^{(j)}_{0:n})} = \frac{g(y_n \mid x_n^{(i)})}{\sum_{j=1}^N g(y_n \mid x_n^{(j)})}. \tag{4.20}$$

Their weighted empirical distribution:

$$\hat{p}(x_{0:n} \mid y_{0:n}) := \frac{1}{N} \sum_{i=1}^{N} W_n^{(i)} \delta_{x_{0:n}^{(i)}}(dx_{0:n}), \tag{4.21}$$

$$\hat{p}(y_n \mid y_{0:n-1}) := \sum_{i=1}^{N} W_{n-1}^{(i)} g(y_n \mid x_n^{(i)}), \tag{4.22}$$

are then approximations of $p(x_{0:n} \mid y_{0:n})$ and $p(y_n \mid y_{0:n-1})$. After correction step, one can resample particles $\left\{x_{0:n}^{(i)}\right\}_{i=1}^{N}$ to obtain N new equally weighted particles $\left\{\tilde{x}_{0:n}^{(i)}, \frac{1}{N}\right\}_{i=1}^{N}$ which construct the associated unweighted empirical distribution $\frac{1}{N}\sum_{i=1}^{N} \delta_{\tilde{x}_{0:n}^{(i)}}(dx_{0:n})$ of $p(x_{0:n} \mid y_{0:n})$.

As we mentioned, applying the bootstrap filter (SMC) to $p(x_{0:n} \mid y_{0:n})$ may suffer from the particle path degeneracy problem. Therefore, in general, the bootstrap filter has been mainly applied to approximate the filtering density $p(x_n \mid y_{0:n})$. In this case, one can derive the algorithm more intuitively by mimicking prediction/filtering recursions of HMMs. Suppose that one has unweighted particle system $\left\{\tilde{x}_{n-1}^{(i)}, \frac{1}{N}\right\}_{i=1}^{N}$ for $p(x_{n-1} \mid y_{n-1})$. Then new particles $\{x_n^{(i)}\}_{i=1}^{N}$ are sampled from $f(x_n \mid \tilde{x}_{n-1}^{(i)})$ which are distributed approximately according to $p(x_n \mid y_{0:n-1})$, from (4.11). That is, their unweighted empirical distribution:

$$\hat{\pi}_{n|n-1}(dx_n) := \frac{1}{N} \sum_{i=1}^{N} \delta_{x_n^{(i)}}(dx_n), \tag{4.23}$$

is approximation of $p(x_n \mid y_{0:n-1})$. Then plugging this measure into the recursion (4.10) gives rise to the following empirical distribution of $p(x_n \mid y_{0:n})$:

$$\begin{aligned} \hat{\pi}_{n|n}(dx_n) &:= \frac{g(y_n \mid x_n)\hat{\pi}_{n|n-1}(dx_n)}{\int g(y_n \mid x_n)\hat{\pi}_{n|n-1}(dx_n)}, \\ &= \frac{\sum_{i=1}^{N} g(y_n \mid x_n^{(i)})\delta_{x_n^{(i)}}(dx_n)}{\sum_{i=1}^{N} g(y_n \mid x_n^{(i)})}. \end{aligned} \tag{4.24}$$

As before, if we set the normalised weights:

$$W_n^{(i)} := \frac{g(y_n \mid x_n^{(i)})}{\sum_{j=1}^{N} g(y_n \mid x_n^{(j)})}, \tag{4.25}$$

then (4.24) becomes:

$$\hat{\pi}_{n|n}(dx_n) = \sum_{i=1}^{N} W_n^{(i)} \delta_{x_n^{(i)}}(dx_n). \tag{4.26}$$

To obtain the "unweighted" empirical distribution as before, we then apply resampling. A resampling procedure associates a number of offspring $N_n^{(i)} \in \mathbb{N}$ with each particle $\{x_n^{(i)}\}_{i=1}^{N}$ with $\sum_{i=1}^{N} N_n^{(i)} = N$.

After resampling, one obtains new particles $\{\tilde{x}_n^{(i)}\}_{i=1}^N$ with associated empirical distribution:

$$\tilde{\pi}_{n|n}(dx_n) := \frac{1}{N}\sum_{i=1}^{N}\delta_{\tilde{x}_n^{(i)}}(dx_n), \tag{4.27}$$

and resulting particles are approximately also distributed according to $p(x_n \mid y_{0:n})$. Next notice that the likelihood function $p(y_{0:n})$ can be decomposed into:

$$p(y_{0:n}) = p(y_0)\prod_{i=1}^{n} p(y_i \mid y_{0:i-1}), \tag{4.28}$$

where $p(y_i \mid y_{0:i-1})$ is given by $p(y_i \mid y_{0:i-1}) = \int g(y_i \mid x_i)p(x_i \mid y_{0:i-1})dx_i$. Therefore, the estimator of the likelihood can then be estimated via the decomposition (4.28) and inserting the empirical predictive distribution (4.27) into $\int g(y_i \mid x_i)p(x_i \mid y_{0:i-1})dx_i$:

$$\hat{p}(y_{0:n}) := \prod_{t=0}^{n}\left[\frac{1}{N}\sum_{i=1}^{N} w_n(x_n^{(i)})\right], \tag{4.29}$$

where $w_n(x_n^{(i)}) := g(y_t \mid x_t^{(i)})$. We algorithmically summarise the above as follows.

Algorithm 14 Bootstrap Filter (Gordon et al., 1993).

Assume that at time $n-1$, one has an equally weighted particle system $\left(\tilde{x}_{n-1}^{(i)}, \frac{1}{N}\right)_{i=1}^N$ of the target $p(x_{n-1} \mid y_{0:n-1})$.

i) Propagate particles $\{x_n^{(i)}\}_{i=1}^N$ via sampling from $f(\cdot \mid \tilde{x}_{n-1}^{(i)})$.

ii) Correct unnormalised weights via $w_n(x_n^{(i)}) = g(y_n \mid x_n^{(i)})$ for $i = 1, \cdots N$.

iii) Obtain $\hat{p}(y_{0:n}) = \prod_{t=0}^{n}\left[\frac{1}{N}\sum_{i=1}^{N} g(y_t \mid x_t^{(i)})\right]$.

iv) Obtain normalised weights via $W_n^{(i)} = \frac{w_n(x_n^{(i)})}{\sum_{j=1}^{N} w_n(x_n^{(j)})}$ for $i = 1, \cdots N$.

v) Do resampling $\left\{x_n^{(i)}\right\}_{i=1}^N$ w.p. $W_n^{(i)}$ to obtain equally weighted particle system $(\tilde{x}_n^{(i)}, \frac{1}{N})_{i=1}^N$.

vi) Return to the first step.

Example 16. *Bootstrap filter for SV model*

Here we apply the bootstrap filter (Algorithm 14) to SV model (Example 14). In this case, we have $f(x' \mid x) = \mathcal{N}(x'; \alpha x, \sigma_x^2)$ and $g(y \mid x) = \mathcal{N}(y; 0, \beta^2 \exp(x))$. As before we set $n = 1000$ and $(\alpha, \sigma_x, \beta) = (0.9, \sqrt{0.1}, 0.8)$. We used $N = 1024$ particles, and did resampling when the effective sample size was less than 512. The results are plotted in Figure 4.

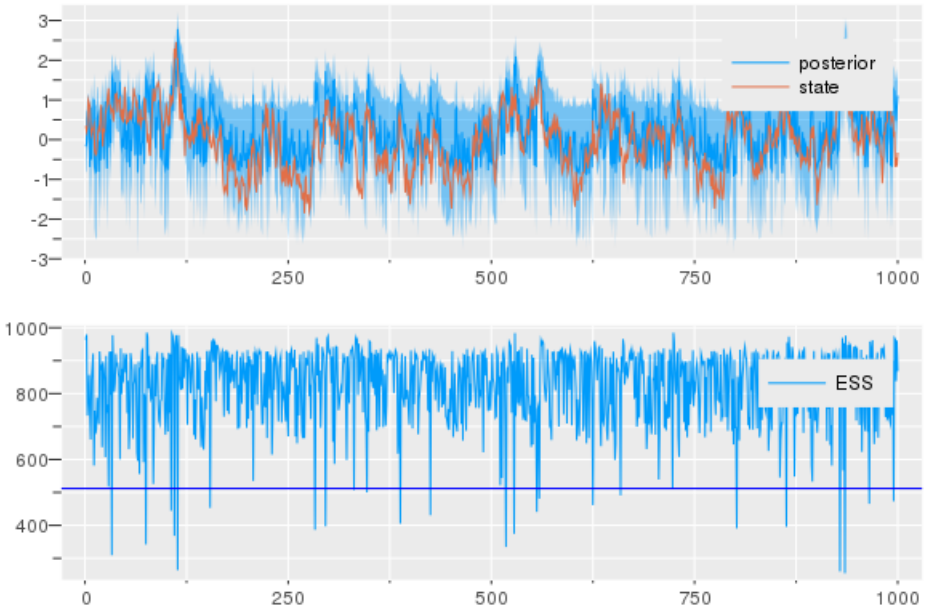


Figure 4: The bootstrap filter (Algorithm 14) estimates for the SV model (Example 14) and the effective sample size. The top figure shows the result of the particle filtering of SV model with the choice $(\alpha, \sigma, \beta) = (0.9, \sqrt{0.1}, 0.8)$ and $N = 1024$ particles. The thick blue line is the estimated posterior mean, area plot is estimated $+/-1$ S.D. and the orange line is the true volatility, that is $\{x_n\}$. The bottom figure shows the effective sample size, we did resampling when the ESS was lower than $1024/2 = 512$.

Next consider the bootstrap filter as a mean field approximation. Again, we want to approximate the sequence of probability measures $(\eta_n)_{n\in\mathbb{N}}$ which correspond to $p(dx_n \mid y_{0:n-1})$. Define $(\mathbb{X}^N, \mathcal{B}(\mathbb{X})^N) := \left(\prod_{p=1}^N \mathbb{X}^p, \prod_{p=1}^N \mathcal{B}(\mathbb{X})^p\right)$. We define the Markov kernel $\mathcal{K}_n(\mathbf{x}_{n-1}, d\mathbf{x}_n)$ from $\mathbb{X}^N$ into $\mathbb{X}^N$ as follows: for any $\mathbf{x}_n \in \left(x_n^{(1)}, \cdots, x_n^{(N)}\right)$, we set:

$$\mathcal{K}_n(\mathbf{x}_{n-1}, d\mathbf{x}_n) := \prod_{p=1}^N \sum_{i=1}^N \frac{g_n(y_{n-1} \mid x_{n-1}^{(i)})}{\sum_{k=1}^N g_n(y_{n-1} \mid x_{n-1}^{(k)})} f_n(dx_n^{(p)} \mid x_{n-1}^{(i)}). \tag{4.30}$$

That is, in order to approximate $p(x_n \mid y_{0:n})$, one can *select* $\mathbf{x}_{n-1}$ with probabilities proportional to $\left\{g_n(y_{n-1} \mid x_{n-1}^{(i)})\right\}_{i=1}^N$. Given this, one can *mutate* the selected vector $\mathbf{x}_{n-1}$ conditionally independently to new positions using the Markov kernel $\mathcal{K}_n(\mathbf{x}_{n-1}, d\mathbf{x}_n)$. Define the empirical measure $\eta_n^N := \frac{1}{N}\sum_{i=1}^N \delta_{x_n^{(i)}}$ on $\mathbb{X}^N$ and recall that $\gamma_{n+1}(1) = \prod_{0\le p\le n} \eta_p(G_p)$ holds. Therefore, we have that $\gamma_{n+1}^N(1) = \prod_{p=0}^n \left[\frac{1}{N}\sum_{i=1}^N g(y_p \mid x_p^{(i)})\right]$. Therefore, the asymptotic results we studied for interacting mean field approximation can be directly applied to the bootstrap filter, see also Crisan and Doucet (2002) for an in-depth treatment of analysis of the particle filter targetting $p(x_n \mid y_{:n})$ from a different approach.

4.4.2 Twisting/Auxiliary Particle Filter

Another important class of particle filter is *Twisting/Auxiliary* particle filter. Assume the now our target is the joint smoothing density $p(x_{0:n} \mid y_{0:n})$ at the final time n only. The optimal (in terms of the conditional variance) choice of the importance density $q^\star(x_n \mid x_{0:n-1})$ is given by $p(x_n \mid x_{n-1}, y_n) = \frac{f(x_n|x_{n-1})g(y_n|x_n)}{p(y_n|x_{n-1})}$ but this is not tractable due to the the predictive likelihood $p(y_n \mid x_{n-1})$ in general. However, this observation implies that a good importance density would take into account information from future observations at the current step. The *Auxiliary Particle Filter* (APF) is a look ahead method in the sense that at time n the algorithm tries to include information from time $n+1$, first studied in Pitt and Shephard (1999) and then generalised in Johansen and Doucet (2008). We follow the later approach. Then the APF can be understood as a standard particle filter applied to the following target distributions:

$$\gamma_n(x_{0:n}) = \hat{p}(x_{0:n}, y_{0:n+1}) := p(x_{0:n}, y_{0:n})\hat{p}(y_{n+1} \mid x_n), \tag{4.31}$$

where $\hat{p}(y_{n+1} \mid x_n)$ is an approximation of the predictive likelihood $p(y_{n+1} \mid x_n) = \int f(x_{n+1} \mid x_n)g(y_{n+1} \mid x_{n+1})dx_{n+1}$. Thus now our target $\pi(x_{0:n})$ is:

$$\pi(x_{0:n}) = \hat{p}(x_{0:n} \mid y_{0:n+1}) := \frac{p(x_{0:n}, y_{0:n})\hat{p}(y_{n+1} \mid x_n)}{\int p(x_{0:n}, y_{0:n})\hat{p}(y_{n+1} \mid x_n)dx_n}. \tag{4.32}$$

Also we set $q(x_n \mid x_{0:n-1}) = q(x_n \mid x_{n-1}, y_n)$ in general. In this setting, the incremental weight is given by:

$$\begin{aligned} a_n(x_{0:n}) &:= \frac{\gamma_n(x_{0:n})}{\gamma_{n-1}(x_{0:n-1})q(x_n \mid x_{0:n-1})} = \frac{p(x_{0:n}, y_{0:n})\hat{p}(y_{n+1} \mid x_n)}{p(x_{0:n-1}, y_{0:n-1})\hat{p}(y_n \mid x_{n-1})q(x_n \mid x_{n-1}, y_n)}, \\ &= \frac{f(x_n \mid x_{n-1})g(y_n \mid x_n)\hat{p}(y_{n+1} \mid x_n)}{\hat{p}(y_n \mid x_{n-1})q(x_n \mid x_{n-1}, y_n)}. \end{aligned} \tag{4.33}$$

Therefore, we can define the associated unnormalised weights:

$$w_n^{APF}(x_{n-1:n}^{(i)}) := a_n(x_{0:n}) = \frac{f(x_n \mid x_{n-1})g(y_n \mid x_n)\hat{p}(y_{n+1} \mid x_n)}{\hat{p}(y_n \mid x_{n-1})q(x_n \mid x_{n-1}, y_n)}. \tag{4.34}$$

Before we summarise, we note that the APF approximates the distributions $\{\hat{p}(x_{0:n} \mid y_{0:n+1})\}$ not $p(x_{0:n} \mid y_{0:n})$. To obtain an empirical representation of $p(x_{0:n} \mid y_{0:n})$, first notice that:

$$p(x_{0:n} \mid y_{0:n}) = p(x_{0:n-1} \mid y_{0:n})p(x_n \mid x_{n-1}, y_n).$$

Therefore one can use importance sampling with:

$$\hat{q}(x_{0:n} \mid y_{0:n}) := \hat{p}(x_{0:n-1} \mid y_{0:n})q(x_n \mid x_{n-1}, y_n). \tag{4.35}$$

as the importance density. Indeed, we can show that for $h \in \mathcal{B}_b(\mathbb{X}^{(n+1)})$:

$$\begin{aligned}
\mathbb{E}_{p(x_{0:n}|y_{0:n})}\left[h(x_{0:n})\right] &= \mathbb{E}_{\hat{q}(x_{0:n}|y_{0:n})}\left[h(x_{0:n})\frac{p(x_{0:n} \mid y_{0:n})}{\hat{q}(x_{0:n} \mid y_{0:n})}\right], \\
&= \mathbb{E}_{\hat{q}(x_{0:n}|y_{0:n})}\left[h(x_{0:n})\frac{p(x_{0:n-1} \mid y_{0:n})p(x_n \mid x_{n-1}, y_n)}{\hat{p}(x_{0:n-1} \mid y_{0:n})q(x_n \mid x_{n-1}, y_n)}\right], \\
&\propto \mathbb{E}_{\hat{q}(x_{0:n}|y_{0:n})}\left[h(x_{0:n})\frac{p(x_{0:n-1} \mid y_{0:n})f(x_n \mid x_{n-1})g(y_n \mid x_n)}{p(x_{0:n-1} \mid y_{0:n-1})\hat{p}(y_{n+1} \mid x_n)q(x_n \mid x_{n-1}, y_n)}\right], \\
&= \mathbb{E}_{\hat{q}(x_{0:n}|y_{0:n})}\left[h(x_{0:n})\frac{f(x_n \mid x_{n-1})g(y_n \mid x_n)}{\hat{p}(y_n \mid x_{n-1})q(x_n \mid x_{n-1}, y_n)}\right]. \qquad (4.36)
\end{aligned}$$

Thus, from (4.36), we define the associated unnormalised importance weight:

$$\tilde{w}_n(x_{n-1:n}^{(i)}) := \frac{f(x_n \mid x_{n-1})g(y_n \mid x_n)}{\hat{p}(y_n \mid x_{n-1})q(x_n \mid x_{n-1}, y_n)}. \qquad (4.37)$$

For the sake of simplicity, we assume that resampling is performed every time steps so that the associated normalised weight is:

$$\tilde{W}_n^{(i)} := \frac{\tilde{w}_n(x_{n-1:n}^{(i)})}{\sum_{j=1}^N \tilde{w}_n(x_{n-1:n}^{(j)})}. \qquad (4.38)$$

As a result, we can construct the following empirical distribution of $p(x_{0:n} \mid y_{0:n})$:

$$\hat{p}(x_{0:n} \mid y_{0:n}) = \sum_{i=1}^N \tilde{W}_n^{(i)} \delta_{x_{0:n}^{(i)}}(dx_{0:n}), \qquad (4.39)$$

and estimator of $p(y_n \mid y_{0:n-1})$ is thus:

$$\hat{p}(y_n \mid y_{0:n-1}) = \left(\frac{1}{N}\sum_{i=1}^N \tilde{w}_n(x_{n-1:n}^{(i)})\right)\left(\tilde{W}_{n-1}^{(i)}\hat{p}(y_n \mid x_{n-1}^{(i)})\right). \qquad (4.40)$$

If it is possible to choose $q(x_n \mid x_{n-1}, y_n) = p(x_n \mid x_{n-1}, y_n)$ and $\hat{p}(y_n \mid x_{n-1}) = p(y_n \mid x_{n-1})$, then the algorithm is called *perfect adaptation*. In this case, the APF takes the simple form as $a_n(x_{0:n}) = p(y_n \mid x_{n-1})$ and $w_n^{APF}(x_{n-1:n}) = 1$. The APF now can be summarised as follows. Notice that the importance density $\hat{q}(x_{0:n} \mid y_{0:n})$ in (4.36) is what is exactly obtained after the propagation step in Algorithm 15 but before the correction step.

Algorithm 15 Auxiliary Particle Filter (Pitt and Shephard, 1999; Johansen and Doucet, 2008).

Assume that at time $n-1$, one has an equally particle system $\left(\tilde{x}_{0:n-1}^{(i)}, \frac{1}{N}\right)_{i=1}^{N}$ of the target $\hat{p}(x_{0:n-1} \mid y_{0:n})$.

i) Propagate particles $\{x_n^{(i)}\}_{i=1}^N$ via sampling from $q(\cdot \mid \tilde{x}_{n-1}^{(i)}, y_n)$ and $x_{0:n}^{(i)} \leftarrow \{x_n^{(i)}, \tilde{x}_{0:n-1}^{(i)}\}_{i=1}^N$.

ii) Correct unnormalised weights via $w_n^{APF}(x_{n-1:n}^{(i)}) = \frac{f(x_n^{(i)} \mid x_{n-1}^{(i)}) g(y_n \mid x_n^{(i)}) \hat{p}(y_{n+1} \mid x_n^{(i)})}{\hat{p}(y_n \mid x_{n-1}^{(i)}) q(x_n^{(i)} \mid x_{n-1}^{(i)}, y_n)}$ for $i = 1, \cdots N$.

iii) Obtain normalised weights via $W_n^{APF(i)} := \frac{w_n^{APF}(x_{n-1:n}^{(i)})}{\sum_{j=1}^N w_n^{APF}(x_{n-1:n}^{(j)})}$ for $i = 1, \cdots N$.

iv) Do resampling $\left\{x_n^{(i)}\right\}_{i=1}^N$ w.p. $W_n^{APF(i)}$ to obtain equally weighted particle system $(\tilde{x}_{0:n}^{(i)}, \frac{1}{N})_{i=1}^N$.

v) Return to the first step.

The main idea of the APF is that one can design target distributions to maximise the accuracy of an approximation of the final target distribution $p(x_{0:n} \mid y_{0:n})$ by changing intermediate target distributions $\{p(x_{0:k} \mid y_{0:k})\}_{1 \leq k < n}$ to take into account information from future observations. In particular, assume that now we are now mainly interested in the likelihood function $p(y_{1:n})$. Given HMMs, then Guarniero et al. (2017) introduce the *twisted hidden Markov models.* First we introduce a sequence of real-valued, bounded, continuous and positive functions $\Psi := (\psi_1, \psi_2, \cdots \psi_n)$. Then we define $f(x, \psi) := \int \psi(x') f(x' \mid x) dx'$ and $\tilde{\psi}_k(x_k) := f(x_k, \psi_{k+1})$ for $k \in \{1, \cdots n-1\}$ with $\tilde{\psi}_n := 1$ and $\tilde{\psi}_0 := \int \eta(x_1)\psi(x_1) dx_1$ where $\eta(x_1)$ is an initial density. Using these, we can define the twisted model as follows:

$$\begin{cases} \eta_1^{\psi_1}(x_1) & := \frac{\eta(x_1)\psi(x_1)}{\tilde{\psi}_0}, \\ f_k^{\psi}(x_k \mid x_{k-1}) & := \frac{f(x_k \mid x_{k-1})\psi_k(x_k)}{\tilde{\psi}_{k-1}(x_{k-1})}, \end{cases} \tag{4.41}$$

for $k \in \{1, \cdots n-1\}$. Also, we define:

$$\begin{cases} g_1^{\psi}(x_1) & := g(x_1 \mid y_1) \frac{\tilde{\psi}_1(x_1)}{\psi_1(x_1)} \tilde{\psi}_0(x_0), \\ g_k^{\psi}(x_k) & := g(x_k \mid y_k) \frac{\tilde{\psi}_k(x_k)}{\psi_k(x_k)}, \end{cases} \tag{4.42}$$

for $k \in \{1, \cdots n-1\}$. Critically, one can show that $Z_\varphi := \int \eta_1^{\psi_1}(x_1) g_1^{\psi}(x_1) \prod_{k=2}^n f_k^{\psi}(x_k \mid x_{k-1}) g_k^{\psi}(x_k) dx_{1:k} = p(y_{1:n})$. To derive the optimal choice of Ψ, consider $\Psi^\star := (\psi_1^\star, \psi_2^\star, \cdots \psi_n^\star)$ such that:

$$\psi_k^\star(x_k) := g(x_k \mid y_k) \mathbb{E}\left[\prod_{p=k+1}^{n} g(x_p \mid y_p) \mid x_p\right], \tag{4.43}$$

$k \in \{1, \cdots n-1\}$. Then it can be shown that particle approximation of $Z_{\varphi^\star}$, say $Z_{\varphi^\star}^N$ is equal to the likelihood function $p(y_{1:n})$ w.p.1, see Guarniero et al. (2017, Proposition 2). Also notice that setting $\psi(x_k) = g(y_k \mid x_k)$ for $k \in \{1, \cdots n-1\}$ gives rise to the fully adapted APF.

In practice, one cannot use the optimal sequence $\Psi^\star$. To evaluate $f(x, \psi)$ and $\tilde{\psi}(x)$ pointwise, one

needs to impose some restrictions on them. First, the initial distribution of HMMs is a a mixture of Gaussians, and the transition densities are of the form:

$$f(\cdot \mid x) = \sum_{k=1}^{N} c_k \mathcal{N}(\cdot; a_k(x), b_k(x)), \tag{4.44}$$

where $\sum_{k=1}^{N} c_k = 1$, $a_k(x)$ and $b_k(x)$ are sequences of mean and covariance functions. Also Ψ defines the class of functions of the form:

$$\psi(x) = C + \sum_{k=1}^{N} c_k \mathcal{N}(x; a_k, b_k), \tag{4.45}$$

where $C \in \mathbb{R}_+$. Under this setting, both $f(x, \psi)$ and $\tilde{\psi}(x)$ can be computed analytically, and $f_k^{\psi}(x_k \mid x_{k-1})$ is a mixture of normal distributions whose component means and covariance matrices can also be computed analytically for $k \in \{1, \cdots n-1\}$. Then to approximate $\Psi^\star$ recursively, Guarniero et al. (2017) consider the following procedure. First observe that $\Psi^\star$ satisfies $\psi_n^\star(x_n) = g(y_n \mid x_n)$ and:

$$\psi_k^\star(x_k) = g(y_k \mid x_k) f(x_k, \psi_{k+1}^\star), \tag{4.46}$$

for $k \in \{1, \cdots n-1\}$ holds, see Guarniero et al. (2017, Proposition 4). In particular, Guarniero et al. (2017) consider the following iterative refinement scheme to tune the parameters (4.46):

$$\begin{cases} (\hat{a}_k, \hat{b}_k, \hat{\lambda}_k) & = \arg\min_{(a,b,\lambda)} \sum_{i=1}^{N} \left[\mathcal{N}(x_n^{(i)}; a, b) - \lambda \psi_k^{(i)} \right]^2, \\ \psi_k(x_k) & := \mathcal{N}\left(x_k; \hat{a}_k, \hat{b}_k\right) + c\left(N, \hat{a}_k, \hat{b}_k\right), \end{cases} \tag{4.47}$$

$k \in \{1, \cdots n-1\}$ with $\psi_n^\star(x_n) = g(y_n \mid x_n)$ where c is a positive real-valued function, see Guarniero et al. (2017, Section 5.1) for details. We write $\rho := (a, b, \lambda)$ in (4.47). Given the updated set of tuning parameters, we then run again particle filter and repeat the update procedure (4.47).

tar

Algorithm 16 Iterated Auxiliary Particle Filter (Guarniero et al., 2017).

i) Initialise $\rho^0 := \left(\rho_1^0, \rho_2^0, \cdots \rho_n^0\right)$.

ii) Given the updated set of tuning parameters ρ^l, run Algorithm 14 for the twisted HMM specified in (4.41), (4.42).

iii) Update tuning parameters ρ^{l+1} using (4.47).

iv) Set $l \leftarrow l + 1$ and back to the second step until user specified threshold is satisfied.

v) Run Algorithm 14 for the twisted HMM again and obtain particle approximation of the likelihood function.

Remark 5. *Algorithm 16 involves a stopping criteria. Guarniero et al. (2017) propose a stopping criteria based on the asymptotic variance of the estimate of the likelihood function.*

4.5 Parameter Inference for HMMs via Particle Filter

In this section, we briefly present statistical inference for HMMs. First we consider a frequentist parameter estimation problem, which relies on the (log) likelihood function:

$$p_\theta(y_{0:n}) := \int \eta_\theta(x_0) g_\theta(y_0 \mid x_0) \prod_{k=1}^{n} g_\theta(y_k \mid x_k) f_\theta(x_k \mid x_{k-1}) dx_{0:n}, \tag{4.48}$$

$$\ell_\theta(y_{0:n}) := \log p_\theta(y_{0:n}) = \sum_{k=0}^{n} \log p_\theta(y_k \mid y_{0:k-1}). \tag{4.49}$$

Then, from a frequentist point of view, the most popular estimator is the maximum likelihood estimator (MLE) which is defined as:

$$\hat{\theta}_n := \arg\max_{\theta \in \Theta} \ell_\theta(y_{0:n}). \tag{4.50}$$

In HMMs context, asymptotic properties such as strong consistency, asymptotic normality of the MLE have been well studied under different conditions. We refer to Douc et al. (2004, 2011b); Douc and Moulines (2012) and Douc et al. (2014, Chapter 13) for this direction. It should be emphasised that, although such analytical results have rather important and insightful meanings, the likelihood function is not analytically available.

In addition to frequentist parameter estimation, we also consider Bayesian inference for HMMs. In a Bayesian framework, one first needs to specify a prior distribution $\pi(\theta)$ on the parameter space Θ. For the sake of simplicity, we assume $\pi(\theta)$ has the density w.r.t. the reference measure $d\theta$ and that is not improper. Then Bayesian inference involves the posterior distribution:

$$\Pi(\theta \mid y_{0:n}) := \frac{p_\theta(y_{0:n})\pi(\theta)}{\int p_\theta(y_{0:n})\pi(\theta) d\theta}. \tag{4.51}$$

In a Bayesian framework, one can make use of the posterior mean and the mode of the posterior distribution as point estimates, defined respectively as:

$$\hat{\theta}_n^{PM} := \int_\Theta \theta \Pi(\theta \mid y_{0:n}) d\theta, \tag{4.52}$$

$$\hat{\theta}_n^{MAP} := \arg\max_{\theta \in \Theta} p_\theta(y_{0:n})\pi(\theta), \tag{4.53}$$

for instance. The posterior consistency implies that as $n \to \infty$, the posterior $\Pi(\theta \mid y_{0:n})$ converges to $\delta_{\theta_\star}$ w.p.1, where $\theta_\star$ denotes the true parameter and $\delta_{\theta_\star}$ denotes the Dirac mass located at the true parameter. This convergence also holds for HMMs (Douc et al., 2016a; Gassiat and Rousseau, 2014). Also well-known Bernstein–von Mises theorem is available for HMMs (De Gunst and Shcherbakova, 2008). Therefore, in this sense, frequentist inference and Bayesian inference are asymptotically equivalent even for HMMs from a frequentist point of view. Again, the posterior distribution is not analytically available for most HMMs.

To overcome the difficulties arising from the intractability, we will resort to particle filter to estimate parameters from both frequentist and Bayesian perspective. In an off-line framework, one infers θ using

fixed observations (batch data) $y_{0:n}$. Thus the computational complexity will increase as n increases. In contrast, on-line methods update the parameter estimate sequentially as observations $\{y_n\}_{n\geq 0}$ become available. Generally, frequentist methods can be reduced to an online setting, and Bayesian ones are off-line since they typically involve MCMC within particle filter.

4.5.1 Frequentist Methods

We tackle the MLE estimation problem for HMMs via particle filter. In particular, following Del Moral et al. (2010) closely, we will focus on two popular computational methods to obtain the MLE, that is the gradient ascent (descent) and Expectation-Maximization (EM) algorithm (Dempster et al., 1977). As for the gradient ascent, one might need to compute the score function $\nabla \log p_\theta(y_{0:n})$. Under mild conditions (Cappé et al., 2005), Fisher's identity enables us to express the score function as:

$$\nabla \log p_\theta(y_{0:n}) = \sum_{k=1}^{n} \mathbb{E}_\theta \left[\nabla \log f_\theta(x_k \mid x_{k-1}) \mid y_{0:n}\right] + \sum_{k=1}^{n} \mathbb{E}_\theta \left[\nabla \log g_\theta(y_k \mid x_{k-1}) \mid y_{0:n}\right]. \tag{4.54}$$

As for the the EM algorithm, the objective function at iteration at $i+1$ (E-step):

$$\begin{aligned} Q(\theta_i, \theta) &:= \int \log p_\theta(x_{0:n}, y_{0:n}) p_{\hat{\theta}_i}(x_{0:n} \mid y_{0:n}) dx_{0:n} \\ &= \sum_{k=1}^{n} \mathbb{E}_{\theta_i} \left[\log f_\theta(x_k \mid x_{k-1})\right] + \sum_{k=0}^{n} \mathbb{E}_{\theta_i} \left[\log g_\theta(y_k \mid x_k) \mid y_{0:n}\right] \end{aligned} \tag{4.55}$$

would be calculated and then maximized. New sequence of parameter estimates $\hat{\theta}_{i+1}$ is obtained as such maximizing argument. Therefore, it turns out that E-step of the EM algorithm and obtaining the score function can be reduced to computation the joint smoothing distribution $p_\theta(x_{0:n} \mid y_{0:n})$.

Let $s_k : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ and set $S_n : \mathbb{X}^n \to \mathbb{R}$ as $S_n := \sum_{k=1}^{n} s_k(x_{k-1}, x_k)$. Critically, observe that both (4.54) and (4.55) can be expressed as:

$$\mathcal{S}_{\theta,n} := \mathbb{E}_\theta \left[S_n(x_{0:n}) \mid y_{0:n}\right]. \tag{4.56}$$

To be precise, $s_k(x_{k-1}, x_k) = \nabla \log f_\theta(x_k \mid x_{k-1}) + \nabla \log g_\theta(y_k \mid x_{k-1})$ for (4.54) and $\log f_\theta(x_k \mid x_{k-1}) + \log g_\theta(y_k \mid x_k)$ for (4.55). Thus, both EM and gradient ascent can be studied as (joint) smoothing problem for additive functional s_k.

Again, in the context of HMMs, estimating such an intractable distribution is generally done by making use of particle filter. We will consider an online version of smoothing via SMC to approximate (4.56), and make it useful for an online version of gradient ascent and EM algorithm for HMMs.

In order to construct the particle approximations of $p_\theta(x_{0:n} \mid y_{0:n})$, one of straightforward methods would be the *path-space* method, studied in Cappe (2009); Dahlhaus and Neddermeyer (2010) for instance. This method can be done as follows. Let $(x_{0:n}^{(i)}, W_n^{(i)})_{i=1}^N$ be particle approximations of $p_\theta(x_{0:n} \mid y_{0:n})$ in the sense that:

$$\hat{p}_\theta(dx_{0:n} \mid y_{0:n}) := \frac{1}{N} \sum_{i=1}^{N} W_n^{(i)} \delta_{x_{0:n}^{(i)}}(dx_{0:n}), \sum_{i=1}^{N} W_n^{(i)} = 1, \tag{4.57}$$

This approximation can be easily obtained by the bootstrap filter. Then substituting (4.57) into (4.56) might yield estimates of $\mathcal{S}_n^\theta$. Although the method is online and the computational cost is $\mathcal{O}(N)$, as pointed out theoretically (Del Moral et al., 2010) and experimentally (Kantas et al., 2015), such path-space method may suffer from the particle path degeneracy problem which is well-known in the SMC literature. Roughly speaking, as $n \to \infty$, the particle approximations of $p_\theta(x_{0:n} \mid y_{0:n})$ obtained by the path-space method may end up with the same ancestral particle due to the successive resampling steps. Therefore, such approximations may collapse as $n \to \infty$.

To overcome this drawback, one of promising alternative would be the Forward Filtering and Backward Smoothing (FFBS) (Doucet et al., 2000).The FFBS relies on the following basic recursion, which is backward in time (Kitagawa, 1987), for $n \geq k$:

$$\begin{aligned} p_\theta(x_k \mid y_{0:n}) &= \int p_\theta(x_k, x_{k+1} \mid y_{0:k}) dx_{k+1}, \\ &= \int p_\theta(x_k \mid x_{k+1}, y_{0:k}) p_\theta(x_{k+1} \mid y_{0:n}) dx_{k+1}, \end{aligned} \tag{4.58}$$

where the backward Markov density $p_\theta(x_k \mid x_{k+1}, y_{0:k})$ is given by:

$$p_\theta(x_k \mid x_{k+1}, y_{0:k}) = \frac{f_\theta(x_{k+1} \mid x_k) p_\theta(x_k \mid y_{0:k})}{\int f_\theta(x_{k+1} \mid x_k) p_\theta(x_k \mid y_{0:k}) dx_k}. \tag{4.59}$$

We emphasise that (4.59) essentially depends on the assumption that $f_\theta(dx_{k+1} \mid x_k)$ admits the density since the backward kernel is time inhomogeneous, hence its expression is still unclear in our setting. Thus, given the particle approximations $(x_k^{(i)}, W_k^{(i)})_{i=1}^N$ of the filtering density $p_\theta(x_k \mid y_{0:k})$, the particle approximations of (4.59) is given by:

$$\hat{p}_\theta(dx_k \mid x_{k+1}, y_{0:k}) = \sum_{i=1}^N \frac{f_\theta(x_{k+1} \mid x_k^{(i)}) W_k^{(i)}}{\sum_{l=1}^N f_\theta(x_{k+1} \mid x_k^{(l)}) W_k^{(l)}} \delta_k(dx_k). \tag{4.60}$$

Now assume that one has the particle approximations $(x_k^{(i)}, W_{k+1|n}^{(i)})_{i=1}^N$ of the marginal smoothing density $p_\theta(x_{k+1} \mid y_{0:n})$. Plugging such approximations and (4.60) into (4.58) results in:

$$\hat{p}_\theta(dx_k \mid y_{0:n}) = \sum_{j=1}^N W_{k+1|n}^{(j)} \sum_{i=1}^N \frac{f_\theta(x_{k+1}^{(j)} \mid x_k^{(i)}) W_k^{(i)}}{\sum_{l=1}^N f_\theta(x_{k+1}^{(j)} \mid x_k^{(l)}) W_k^{(l)}} \delta_k(dx_k). \tag{4.61}$$

Then the fact that $p_\theta(x_{0:n} \mid y_{0:n}) = p_\theta(x_n \mid y_{0:n}) \prod_{t=0}^{n-1} p_\theta(x_t \mid x_{t+1:t}, y_{0:t})$ and (4.61) will yield the FFBS approximation of (4.56). Apparently this approximation is not online. Also, we note that the computational cost of the FFBS approximation of the for general test functions is $\mathcal{O}(N^t)$ so it is not practical. However, as we will study, the cost of the FFBS can be reduced for the additive functional case.

To reduce an online version of the FFBS for the additive functional, we first define:

$$T_{\theta,n}(x_n) := \int S_n(x_{0:n}) p_\theta(x_{0:n-1} \mid y_{0:n-1}, x_n) dx_{0:n-1}. \tag{4.62}$$

Then it is straightforward to observe that:

$$\begin{aligned}\mathcal{S}_{\theta,n} &= \int\int S_n(x_{0:n}) p_\theta(x_{0:n-1} \mid y_{0:n-1}, x_n) dx_{0:n-1} p_\theta(x_n \mid y_{0:n}) dx_n, \\ &= \int T_{\theta,n}(x_n) p_\theta(x_n \mid y_{0:n}) dx_n. \end{aligned} \tag{4.63}$$

Critically, for any $n \geq k$,

$$T_{\theta,n}(x_n) = \int \left[T_{\theta,n-1}(x_{n-1}) + s_n(x_{n-1}, x_n)\right] p_\theta(x_{n-1} \mid y_{0:n-1}, x_n) dx_{n-1}, \tag{4.64}$$

holds, see Del Moral et al. (2010, Proposition 2.1). Hence, (4.64) and (4.63) give rise to the following online recursion:

$$\mathcal{S}_{\theta,n} = \int \left(\int \left[T_{\theta,n-1}(x_{n-1}) + s_n(x_{n-1}, x_n)\right] p_\theta(x_{n-1} \mid y_{0:n-1}, x_n) dx_{n-1} \right) p_\theta(x_n \mid y_{0:n}) dx_n, \tag{4.65}$$

where the expression of $p_\theta(x_{n-1} \mid y_{0:n-1}, x_n)$ is in (4.59). Therefore, plugging the particle approximations of the backward transition density given in (4.61) and of the filtering density $p_\theta(x_n \mid y_{0:n})$ which is easily obtained via the bootstrap filter, for instance, into (4.65) will give the following forward only implementation of the FFBS for (4.56) with the computational cost $\mathcal{O}(N^2)$.

Algorithm 17 Forward only particle smoothing for HMMs with the additive functionals (Del Moral et al., 2010).

---

i) Assume that at time $n-1$, one has the particle approximations $(x_{n-1}^{(i)}, W_{n-1}^{(i)})_{i=1}^N$ of $p_\theta(x_{n-1} \mid y_{0:n-1})$ and $\left\{\hat{T}_{\theta,n-1}(x_{n-1}^{(i)})\right\}_{i=1}^N$ of $T_{\theta,n-1}(x_{n-1})$.

ii) At time n, sample $x_n^{(i)}$ for $i \cdots N$ from the mixture density (Gordon et al., 1993):

$$x_n^{(i)} \sim \frac{\sum_{j=1}^n f_\theta(x_n \mid x_{n-1}^{(j)}) g_\theta(y_{n-1} \mid x_{n-1}^{(j)})}{\sum_{j=1}^n g_\theta(y_{n-1} \mid x_{n-1}^{(j)})}. \tag{4.66}$$

iii) Then set, for $i \cdots N$:

$$\hat{T}_{\theta,n}(x_n^{(i)}) = \frac{\sum_{j=1}^N W_{n-1}^{(j)} f_\theta(x_n^{(i)} \mid x_n^{(j)})}{\sum_{l=1}^N W_{n-1}^{(l)} f_\theta(x_n^{(i)} \mid x_n^{(l)})} \left[\hat{T}_{\theta,n-1}(x_{n-1}^{(j)}) + s_n(x_{n-1}^{(j)}, x_n^{(i)})\right]. \tag{4.67}$$

iv) Obtain estimate of $\mathcal{S}_{\theta,n}$ as:

$$\hat{\mathcal{S}}_{\theta,n} = \sum_{i=1}^N W_n^{(i)} \hat{T}_{\theta,n}(x_n^{(i)}). \tag{4.68}$$

Poyiadjis et al. (2011) use the score function estimation methodology to propose an *online* gradient ascent algorithm for obtaining an MLE-type parameter estimate, following ideas in LeGland and Mevel

(1997). In more detail, the method is based on the following Robbins–Monro type of recursion:

$$\theta_{n+1} = \theta_n + \gamma_{n+1} \nabla \log p_{\theta_n}(y_n \mid y_{0:n-1}), \tag{4.69}$$

where $\{\gamma_k\}_k$ is a positive decreasing sequence with:

$$\sum_{k=1}^{\infty} \gamma_k = \infty, \sum_{k=1}^{\infty} \gamma_k^2 < \infty.$$

See LeGland and Mevel (1997); Tadic and Doucet (2018) for analytical studies on the convergence properties of the algorithm. In particular, under strict conditions, and cases or trivial models, the algorithm is shown to converge to $\theta_\star$. Note that $\nabla \log p_{\theta_n}(y_n \mid y_{0:n-1}) = \nabla \log p_{\theta_n}(y_{0:n}) - \nabla \log p_{\theta_{n-1}}(y_{0:n-1})$. Therefore, the following online gradient ascent would be directly established by combining (4.69) with the forward particle smoothing we have presented.

Algorithm 18 Online gradient ascent for HMMs via forward particle smoothing (Poyiadjis et al., 2011).

---

i) Assume that at time $n-1$, one has the particle approximations $(x_{n-1}^{(i)}, W_{n-1}^{(i)})_{i=1}^N$ of $p_{\hat{\theta}_{n-1}}(x_{n-1} \mid y_{0:n-1})$ and $\left\{\hat{T}_{\hat{\theta}_{n-1},n-1}(x_{n-1}^{(i)})\right\}_{i=1}^N$ of $T_{\hat{\theta}_{n-1},n-1}(x_{n-1})$.

ii) Update via:

$$\hat{\theta}_n = \hat{\theta}_{n-1} + \gamma_n \left(\hat{\mathcal{S}}_{\hat{\theta}_{n-1},n-1} - \hat{\mathcal{S}}_{\hat{\theta}_{n-2},n-2}\right).$$

iii) At time n, sample $x_n^{(i)}$ for $i = 1 \cdots N$ from the mixture density:

$$x_n^{(i)} \sim \frac{\sum_{j=1}^n f_{\hat{\theta}_n}(x_n^{(i)} \mid x_{n-1}^{(j)}) g_{\hat{\theta}_n}(y_{n-1} \mid x_{n-1}^{(j)})}{\sum_{j=1}^n g_{\hat{\theta}_n}(y_{n-1} \mid x_{n-1}^{(j)})}.$$

iv) Then set:

$$\hat{T}_{\hat{\theta}_n,n}(x_n^{(i)}) = \frac{\sum_{j=1}^N W_{n-1}^{(j)} f_{\hat{\theta}_n}(x_n^{(i)} \mid x_{n-1}^{(j)})}{\sum_{l=1}^N W_{n-1}^{(l)} f_{\hat{\theta}_n}(x_n^{(i)} \mid x_{n-1}^{(l)})} \left[\nabla \log g_{\hat{\theta}_n}(y_n | x_n^{(i)}) + \nabla \log f_{\hat{\theta}_n}(x_n^{(i)} \mid x_{n-1}^{(j)}) + T_{\hat{\theta}_{n-1},n-1}(x_{n-1}^{(j)})\right].$$

v) Obtain estimate of the score function $\nabla \log p_{\hat{\theta}_n}(y_{0:n})$ as:

$$\hat{\mathcal{S}}_{\hat{\theta}_n,n} = \sum_{i=1}^N W_n^{(i)} \hat{T}_{\hat{\theta}_n,n}(x_n^{(i)}).$$

Example 17. *Online gradient ascent via forward particle smoothing for SV model*

Again consider SV model (Example 14). We simulated the data with $n = 50,000$ and $(\alpha, \sigma_x, \beta) = (0.9, \sqrt{0.1}, 0.8)$ as the true parameters. We then applied Algorithm 18 to the simulated data with $N = 150$ particles. The results are plotted in Figure 5.

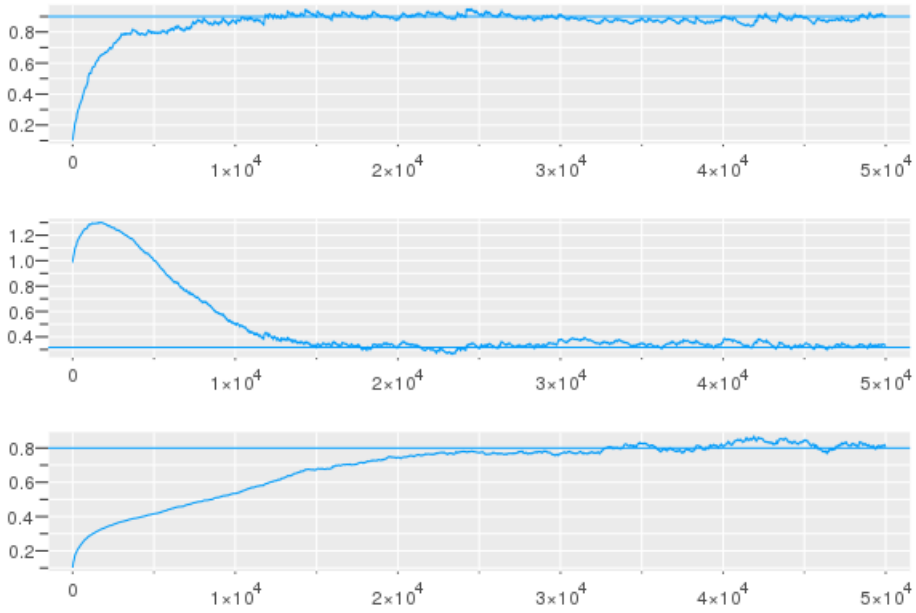


Figure 5: Online estimation of α (top), σ_x (middle) and β (bottom) for the data set simulated according to SV model. We set $(0.1, 1.0, 0.1)$ as the initial values for $(\alpha, \sigma_x, \beta)$ respectively with $N = 150$ particles in Algorithm 18. The horizontal dash lines indicate the true parameter values in each case.

Next, assume that $p_\theta(x_{0:n}, y_{0:n})$ belongs to the exponential family. Let $s^{(l)} : \mathbb{X} \times \mathbb{X} \times \mathbb{Y} \to \mathbb{R}$, $l = 1, 2, \ldots m$ be a collection of functions with corresponding additive functional:

$$S_n^{(l)}(x_{0:n}) := \sum_{k=1}^{n} s^{(l)}(x_{k-1}, x_k, y_k),\ 1 \le l \le m.$$

In this setting, (4.56) for l may become:

$$\mathcal{S}_{n,\theta}^{(l)} = \int S_n^{(l)}(x_{0:n}) p_\theta(x_{0:n}, y_{0:n}) dx_{0:n}. \tag{4.70}$$

Assume that now one has the estimate θ_i of the parameter θ at the iteration step i. Under the assumption that $p_\theta(x_{0:n}, y_{0:n})$ belongs to the exponential family, M-step of the EM algorithm then may be reduced to the following simple iteration:

$$\theta_{i+1} = \Lambda(n^{-1}\mathcal{S}_n(\theta_i)),$$

where $\Lambda : \mathbb{R}^m \to \Theta$ is a suitable function and $\mathcal{S}_{n,\theta}$ is a vector whose l-th component is $\mathcal{S}_{n,\theta}^{(l)}$. To reduce a online EM algorithm via particle smoothing, consider the following average of the sufficient statistic:

$$S_{n+1}^{(l)}(x_{0:n+1}) = \gamma_{n+1} S_{n+1}^{(l)}(x_n, x_{n+1}, y_{n+1}) + (1 - \gamma_n) S_n^{(l)}(x_{0:n}),$$

where $\{\gamma_k\}_{k=1}^{\infty}$ is again a positive decreasing sequence with:

$$\sum_{k=1}^{\infty}\gamma_k = \infty, \sum_{k=1}^{\infty}\gamma_k^2 < \infty.$$

When $\gamma_n = n^{-1}$, important choices are (Del Moral et al., 2010):

$$\gamma_n = n^{-\alpha},\ 0.5 < \alpha \leq 1.$$

Example 18. *The summary statistics for SV model*

Consider SV model (Example 14). The model parameters $(\alpha, \sigma_x^2, \beta) \in (-1,1)\times(0,\infty)\times(0,\infty)$ then are to be estimated. Set $s^{(1)}(x_{n-1}, x_n, y_n) = x_{n-1}x_n$, $s^{(2)}(x_{n-1}, x_n, y_n) = x_{n-1}^2$, $s^{(3)}(x_{n-1}, x_n, y_n) = x_n^3$ and $s^{(4)}(x_{n-1}, x_n, y_n) = y_n^2\exp(-x_n)$. Then M-step of the EM algorithm is characterised by the function such that $\Lambda(z_1, z_2, z_3, z_4) = \left(\frac{z_1}{z_2}, z_3 - \frac{z_1^2}{z_2}, z_4\right)$, see Del Moral et al. (2010) for the details.

Algorithm 19 Online EM for HMMs via forward particle smoothing (Del Moral et al., 2010).

---

i) Assume that at time $n-1$, one has the particle approximates $(x_{n-1}^{(i)}, W_{n-1}^{(i)})_{i=1}^N$ of $p_{\hat{\theta}_{n-1}}(x_{n-1} \mid y_{0:n-1})$ and $\left\{\hat{T}_{\hat{\theta}_{n-1},n-1}(x_{n-1}^{(i)})\right\}_{i=1}^N$ of $T_{\hat{\theta}_{n-1},n-1}(x_{n-1})$.

ii) Update via:

$$\hat{\theta}_n = \Lambda(\hat{\mathcal{S}}_{n-1,\hat{\theta}_{n-1}}).$$

iii) At time n, sample $x_n^{(i)}$ for $i = 1\cdots N$ from the mixture density:

$$x_n^{(i)} \sim \frac{\sum_{j=1}^n f_{\hat{\theta}_n}(x_n^{(i)} \mid x_{n-1}^{(j)}) g_{\hat{\theta}_n}(y_{n-1} \mid x_{n-1}^{(j)})}{\sum_{j=1}^n g_{\hat{\theta}_n}(y_{n-1} \mid x_{n-1}^{(j)})}.$$

iv) Then set:

$$\hat{T}_{\hat{\theta}_n,n}(x_n^{(i)}) = \frac{\sum_{j=1}^N W_{n-1}^{(j)} f_{\hat{\theta}_n}(x_n^{(i)} \mid x_{n-1}^{(j)})}{\sum_{l=1}^N W_{n-1}^{(l)} f_{\theta_{\hat{\theta}_n}}(x_n^{(i)} \mid x_{n-1}^{(l)})}\left[(1-\gamma_n)\hat{T}_{\hat{\theta}_{n-1},n-1}(x_{n-1}^{(j)}) + \gamma_n s(x_{n-1}^{(j)}, x_n^{(i)})\right].$$

v) Obtain estimate of $\mathcal{S}_{n,\theta}$ as:

$$\hat{\mathcal{S}}_{\hat{\theta}_n,n} = \sum_{i=1}^N W_n^{(i)}\hat{T}_{\hat{\theta}_n,n}(x_n^{(i)}).$$

4.5.2 Bayesian Methods

As for Bayesian inference via SMC, We will introduce two particle based methods, particle MCMC (PMCMC) (Andrieu et al., 2010) and SMC^2 (Chopin et al., 2013). First, it should be emphasised that both methods are based fully or partly on the pseudo-marginal MCMC approach Andrieu and Roberts

(2009). To complete Bayesian inference, one has to specify a prior distribution $\pi(d\theta)$ on the parameter space Θ. Without loss of generality, we assume that $\pi(d\theta)$ has the density w.r.t. the Lebesgue measure denoted by $d\theta$ and which is also well defined (not improper prior). In this case, the target will be:

$$p_\theta(x_{0:n} \mid y_{0:n}) = p(\theta, x_{0:n} \mid y_{0:n}).$$

A popular way to estimate such a posterior will be to use MCMC within particle filter. Note that MCMC algorithms often struggle to update the path $x_{0:n}$ if a model has a strong dependency. Also, a strong correlation between θ and $x_{0:n}$ will deteriorate the speed of mixing. These imply that inference based on the joint posterior $p(\theta, x_{0:n} \mid y_{0:n})$ will end up with poor mixing. Therefore, it might be desirable if one could do sampling from:

$$\begin{aligned} \Pi(\theta \mid y_{0:n}) &\propto p(y_{0:n} \mid \theta)\pi(\theta) \\ &\propto \pi(\theta) \int p(x_{0:n} \mid \theta) p(y_{0:n} \mid x_{0:n}, \theta) dx_{0:n}, \end{aligned} \tag{4.71}$$

that is, instead of sampling on the joint space $\mathbb{X}^{(n+1)} \times \Theta$, we want to do it only on the parameter space Θ but again this integral cannot be evaluated analytically in the case of HMMs, in general. All in all, we would like to integrate out $x_{0:n}$ and do sampling from the marginal distribution $\Pi(\theta \mid y_{0:n})$.

To facilitate study, assume that one can obtain $\Pi(\theta \mid y_{0:n})$ analytically. Here, we will focus on Metropolis-Hastings algorithm. Let $q(\theta' \mid \theta)$ denote the proposal density for the parameter θ. Then, at time n, a proposed new value θ' will be accepted according to:

$$\alpha_{MH}(\theta, \theta') := \min\left\{1, \frac{q(\theta \mid \theta')\pi(\theta')p_{\theta'}(y_{0:n})}{q(\theta' \mid \theta)\pi(\theta)p_\theta(y_{0:n})}\right\}. \tag{4.72}$$

Then the Markov chain generated in such way leaves $\Pi(\theta \mid y_{0:n})$ invariant as we studied. Then the critical observation is that, as we studied, a by-product of the particle filter output (4.29):

$$\hat{p}_\theta(y_{0:n}) = \prod_{t=0}^{n} \left[\frac{1}{N}\sum_{i=1}^{N} g(y_t \mid x_t^{(i)})\right],$$

is an unbiased estimate of $p_\theta(y_{0:n})$, see also Del Moral (2004, Proposition 7.4.1) for instance. Therefore $\pi(\theta)\hat{p}_\theta(y_{0:n})$ is a point-wise unbiased estimate of $p(\theta, y_{0:n})$. Then again Andrieu and Roberts (2009) show that if one replaces $p_\theta(y_{0:n})$ by its unbiased estimate $\hat{p}_\theta(y_{0:n})$, MCMC outputs still leaves marginally the target distribution invariant. That is, one can replace the ratio by:

$$\alpha_{PMH}(\theta, \theta') := \min\left\{1, \frac{q(\theta \mid \theta')\pi(\theta')\hat{p}_{\theta'}(y_{0:n})}{q(\theta' \mid \theta)\pi(\theta)\hat{p}_\theta(y_{0:n})}\right\}. \tag{4.73}$$

Indeed, let $u \in \mathcal{U}$ denote the random variable used in the particle filter to construct $\hat{p}_\theta(y_{0:n} \mid u)$, here we use u to emphasise the dependency of a particle estimate of $p_\theta(y_{0:n})$ on u. Indeed, the particle filter is a deterministic given u. Therefore, *the random variables u are equivalent to an estimate of the*

filtering distribution in the context. Then we can define the extended target on $(\Theta, \mathcal{U})$ as:

$$\hat{p}(\theta, u \mid y_{0:n}) \propto \hat{p}_\theta(y_{0:n} \mid u) p(u \mid \theta) \pi(\theta). \tag{4.74}$$

It is clear to see that $\mathbb{E}_u\left[\hat{p}(\theta, u \mid y_{0:n})\right] \propto \pi(\theta) \int \hat{p}_\theta(y_{0:n} \mid u) p(u \mid \theta) du = \pi(\theta) p_\theta(y_{0:n}) \propto p(\theta \mid y_{0:n})$ as required, since $\hat{p}_\theta(y_{0:n} \mid u)$ is unbiased. Also, we have that $\hat{p}(\theta, u \mid y_{0:n}) q(\theta' \mid \theta) p(u' \mid \theta') \alpha_{PMH}(\theta, \theta') = p(u \mid \theta) p(u' \mid \theta') \times \min\left\{q(\theta' \mid \theta) \hat{p}_\theta(y_{0:n} \mid u) \pi(\theta), q(\theta \mid \theta') \hat{p}_{\theta'}(y_{0:n} \mid u') \pi(\theta')\right\}$, so detailed balance holds w.r.t. the extended target $\hat{p}(\theta, u \mid y_{0:n})$ since this is symmetric. Thus, this observation leads to the following algorithm, called *Particle marginal Metropolis-Hastings* (PMMH).

In practice, one does not need to save u for each iteration, but it suffices to save the scalar value for estimate $\hat{p}(\theta, u \mid y_{0:n})$. Note that Andrieu et al. (2010) find that there is a lot more flexibility to use particle filter within MCMC, such as particle Gibbs (conditional particle filter), and develop a number of *Particle MCMC* (PMCMC) algorithms. PMMH is one of PMCMC algorithms. Again, this method is not an on-line method, but an off-line (batch) one. Also, PMCMC methods do not provide tools for sequential analysis.

Algorithm 20 Particle marginal Metropolis-Hastings (PMMH) (Andrieu et al., 2010)

Assume that at the iteration step $i-1$, one has $\hat{\theta}_{i-1}$. Then iterate the followings for $i = 1...M$.

i) Draw θ' from $q(\cdot \mid \hat{\theta}_{i-1})$.

ii) Given θ', run particle filter to obtain $\hat{p}_{\theta'}(y_{0:n})$.

iii) Draw a from $Unif(0,1)$.

iv) Calculate the ratio $\alpha_{PMH}(\hat{\theta}_{i-1}, \theta') = \min\left\{1, \frac{q(\hat{\theta}_{i-1} \mid \theta') \pi(\theta') \hat{p}_{\theta'}(y_{0:n})}{q(\theta' \mid \hat{\theta}_{i-1}) \pi(\hat{\theta}_{i-1}) \hat{p}_{\hat{\theta}_{i-1}}(y_{0:n})}\right\}$.

v) If $a \leq \alpha_i$, accept θ' and $\hat{p}_{\theta'}(y_{0:n})$ and set $\hat{\theta}_i = \theta'$, $\hat{p}_{\hat{\theta}_i}(y_{0:n}) = \hat{p}_{\theta'}(y_{0:n})$. Otherwise, set $\hat{\theta}_i = \hat{\theta}_{i-1}$, $\hat{p}_{\hat{\theta}_i}(y_{0:n}) = \hat{p}_{\hat{\theta}_{i-1}}(y_{0:n})$.

vi) Return to the first step.

Remark 6. *It is also possible to update jointly path $x_{0:n}$ and parameters θ if one is also interested in the joint posterior $p(\theta, x_{0:n} \mid y_{0:n})$. In this case, one also needs to sample a single path $x'_{0:n}$, which is also constructed by $u \in \mathcal{U}$, by running the particle filter. Then accept $x'_{0:n}$, θ' and $\hat{p}_{\theta'}(y_{0:n})$ w.p. $\alpha_{PMH}(\hat{\theta}_{i-1}, \theta')$ in Algorithm 20.*

Example 19. *PIHM for SV model.*

In this example, we apply PMHM (Algorithm 20) to SV model (Example 14). First we simulated data with $n = 250$ and $(\alpha, \sigma_x, \beta) = (0.9, \sqrt{0.1}, 0.8)$. We used $N = 1024$ particles and $M = 10000$ MCMC iterations. As for priors, we used $\mathcal{TN}_{[-1,1]}\left(0.95, 0.05^2\right)$ for α, $\mathcal{TN}_{[0,1]}\left(0.3, 0.05^2\right)$ for σ_x and $\mathcal{TN}_{[0,1]}\left(0.85, 0.05^2\right)$ for β where $\mathcal{TN}_{[a,b]}\left(\mu, \sigma^2\right)$ denotes the truncated Gaussian distribution with mean μ, standard deviation σ in the interval $[a, b]$. The results are plotted in Figure 6 and Figure 7.

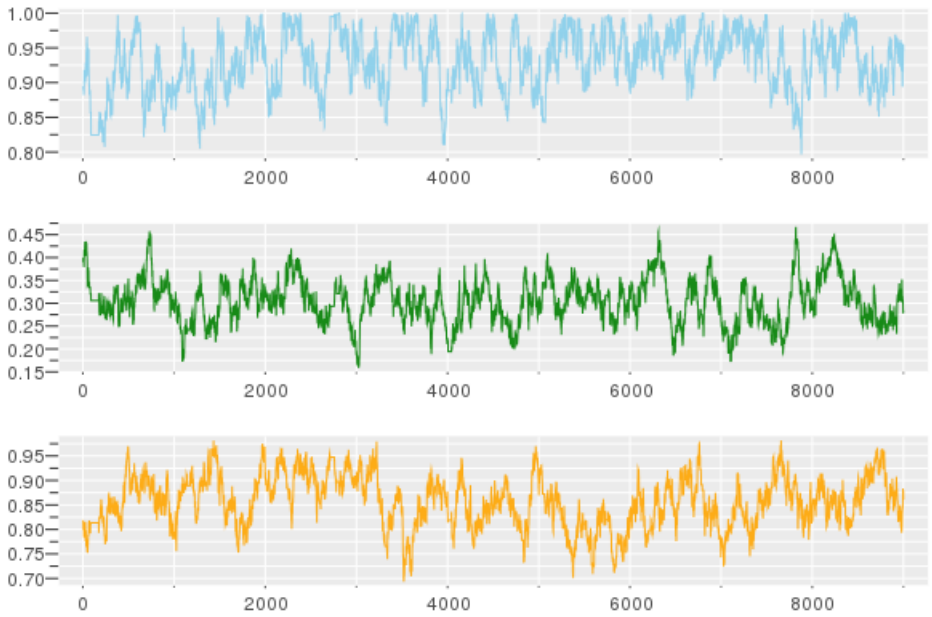


Figure 6: The state of the Markov chain at 1000 iterations after the burn-in for SV model with $(\alpha, \sigma_x, \beta) = (0.9, \sqrt{0.1}, 0.8)$. We used $N = 1024$ particles and $M = 10,000$ MCMC iterations. The blue lines stands for α, the green one stands for σ_x and the orange one stands for β.

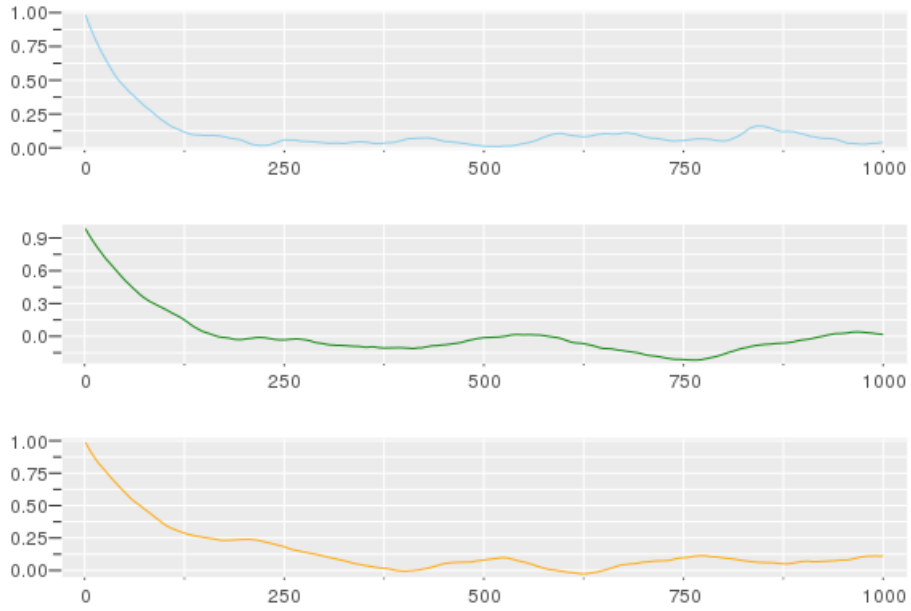


Figure 7: The estimated autocorrelation function of the Markov chain at $1,000$ iterations after the burn-in for SV model with $(\alpha, \sigma_x, \beta) = (0.9, \sqrt{0.1}, 0.8)$. We used $N = 1024$ particles and $M = 10,000$ MCMC iterations. The blue lines stands for α, the green one stands for σ_x and the orange one stands for β.

One of the drawbacks of PMCMC is that, as we mentioned, it cannot be used for sequential analysis. That is, PMCMC does not provide quantities such as $\hat{p}_{\hat{\theta}}(y_n \mid y_{0:n-1})$. Alternatively, SMC^2 (Chopin et al., 2013) provides ways to analyse HMMs sequentially. In particular, SMC^2 is quite useful for model evaluation of HMMs, from a Bayesian point of view. SMC^2 can be understood as a natural amalgamation of particle filter and *Iterated Batch Importance Sampling* (IBIS) (Chopin, 2002), and thus we first begin with a brief explanation of IBIS which is a generalisation of AIS (Neal, 2001).

Assume that now one is interested in a partial posterior $\Pi(\theta \mid y_{0:t})$ $t < n$ and, critically, the predictive likelihood function $p_\theta(y_t \mid y_{0:t-1})$ is, for any t, analytically available. Assume also that now new k observations are available, and $\Pi(\theta \mid y_{0:t})$ and $\Pi(\theta \mid y_{0:t+k})$ are likely to be similar. Then the incremental weights $\alpha(\theta)$ might be given by $\alpha_{t:k}(\theta) \propto \frac{\Pi(\theta|y_{0:t+k})}{\Pi(\theta|y_{0:t})} \propto \frac{p_\theta(y_{0:t+k})}{p_\theta(y_{0:t})} = p_\theta(y_{t+1:t+k} \mid y_{0:t})$. To propagate θ, IBIS involves a Markov kernel $\mathcal{K}_{t+k}(\theta, d\theta')$ which leaves $\Pi(\theta \mid y_{0:t+k})$ invariant. As a special case, we can do sequential analysis by setting $k = 1$ so that we have $\alpha_t(\theta) \propto p_\theta(y_t \mid y_{0:t-1})$. IBIS is then a special case of the particle filter whose target is a partial posterior $\Pi(\theta \mid y_{0:t})$ with the incremental weights $p_\theta(y_t \mid y_{0:t-1})$ and the importance distribution $\mathcal{K}_t(\theta, d\theta')$. We summarise IBIS for sequential analysis as follows, here we assume that resampling occurs each time step. Also, notice that one can construct a consistent estimator of the predictive densities $L_t := \int p_\theta(y_t \mid y_{0:t-1})\pi(\theta)d\theta$.

Algorithm 21 Iterated Batch Importance Sampling (IBIS) (Chopin, 2002)

At time $t = 0$, draw $\{\tilde{\theta}^{(i)}\}_{i=1}^{N_\theta}$ from a prior distribution $\pi(d\theta)$ and set $\{\omega_{0,\tilde{\theta}^{(i)}}\}_{i=1}^{N_\theta} = 1$. Then for $t = 1, \cdots n$, iterate followings.

i) Propagate $\{\theta^{(i)}\}_{i=1}^{N_\theta}$ according to $\Pi(\theta \mid y_{0:t})-$invariant Markov kernel $\mathcal{K}_t(\tilde{\theta}^{(i)}, \cdot)$.

ii) Correct unnormalised weights via $\omega_{t,\theta^{(i)}} := p_{\theta^{(i)}}(y_t \mid y_{0:t-1})$ for $i = 1, \cdots N_\theta$.

iii) Obtain consistent estimate of the predictive density as $\hat{L}_t := \frac{1}{N}\sum_{i=1}^{N} w_{t,\theta^{(i)}}$.

iv) Obtain normalised weights via $\Omega_t^{(i)} := \frac{\omega_{t,\theta^{(i)}}}{\sum_{j=1}^{N}\omega_{t,\theta^{(j)}}}$ for $i = 1, \cdots N_\theta$.

v) Do resampling $\left\{\theta^{(i)}\right\}_{i=1}^{N_\theta}$ w.p. $\Omega_t^{(i)}$ to obtain equally weighted particle system $(\tilde{\theta}^{(i)}, \frac{1}{N_\theta})_{i=1}^{N_\theta}$.

vi) Return to the first step.

As for general HMMs, one cannot evaluate $p_\theta(y_t \mid y_{0:t-1})$ again, and the particle filter provides estimates of it. To be precise, SMC^2 associates N_x $x-$particles to each of the N_θ $\theta-$particles. Also, in order to rejuvenate $\theta-$particles, PMCMC steps are required. We note that whilst PMCMC replaces the likelihood $p_\theta(y_{0:n})$ by particle filter estimates, SMC^2 replaces the incremental likelihood $p_\theta(y_n \mid y_{0:n-1})$ by particle estimates, and thus SMC^2 can sample from $(x_{0:t}, \theta \mid y_{0:t})$ for any $t \in [0, n]$ sequentially. Although Chopin et al. (2013) argue that particle filter estimates of $p_\theta(y_n \mid y_{0:n-1})$ are unbiased, this is not true. The by-product of the SMC outputs is unbiased, as we mentioned. Also, SMC^2 provides the *model evidence* $p(y_{0:n}) = \sum_{t=0}^{n}\int p_\theta(y_t \mid y_{0:t-1})\pi(\theta)d\theta$ which is of particular importance in a Bayesian model selection. We summarise SMC^2 as follows, again we assume that resampling occurs each time step.

Algorithm 22 SMC^2 (Chopin et al., 2013)

At time $t = 0$, draw $\{\theta^{(i)}\}_{i=1}^{N_\theta}$ from a prior distribution $\pi(d\theta)$ and set $\{w_{t,\theta^{(i)}}(x_t)\}_{i=1}^{N_\theta} = 1$. Then for $t = 1, \cdots n$, iterate followings.

i) For each $\{\theta^{(i)}\}_{i=1}^{N_\theta}$, run the particle filter with N_x-particles and obtain $\left\{x_{0:t}^{(i,j)}, \frac{1}{N_x}\right\}_{j=1}^{N_x}$ for $i = 1, \cdots N_\theta$.

ii) Compute particle estimates of $p_{\theta^{(i)}}(y_t \mid y_{0:t-1})$ via $w_{t,\theta^{(i)}}(x_t) := \frac{1}{N_x}\sum_{j=1}^{N_x} g_{\theta^{(i)}}(y_t \mid x_t^{(i,j)})$ for $i = 1, \cdots N_\theta$.

iii) Obtain consistent estimate of the predictive density as $\hat{L}_t := \frac{1}{N}\sum_{i=1}^{N} w_{t,\theta^{(i)}}(x_t)$.

iv) Do sampling $\left\{\tilde{\theta}^{(i)}, \tilde{x}_{0:t}^{(i,1:N_x)}\right\}_{i=1}^{N_\theta}$ via PMCMC and set $(\theta^{(i)}, x_{0:t}^{(i,1:N_x)})_{i=1}^{N_\theta} = (\tilde{\theta}^{(i)}, \tilde{x}_{0:t}^{(i,1:N_x)})_{i=1}^{N_\theta}$.

v) Return to the first step.

Remark 7. *On-line methods require certain iterations to be converged. Since the number of iterations is equal to the one of the data in on-line settings, such on-line methods will need relatively a large sample size. In contrast, the convergence property of the off-line methods which we introduced does not depend on sample size, whilst they are computationally expensive compared with on-line methods. Therefore, on-line methods should be applied to a case where one would be able to access a large sample size. Otherwise, off-line methods might be desirable.*

5 Asymptotic Analysis of Model Selection Criteria for General Hidden Markov Models

5.1 Introduction

The section obtains analytical results for the asymptotic properties of Model Selection Criteria – widely used in practice – for a general family of hidden Markov models (HMMs), thereby substantially extending the related theory beyond typical 'i.i.d.-like' model structures and filling in an important gap in the relevant literature. In particular, we look at the Bayesian and Akaike Information Criteria (BIC and AIC) and the model evidence. In the setting of nested classes of models, we prove that BIC and the evidence are strongly consistent for HMMs (under regularity conditions), whereas AIC is not weakly consistent. Numerical experiments support our theoretical results.

Model Selection has been one of the most well studied topics in Statistics. BIC (Schwarz, 1978) or AIC (Akaike, 1974) - as well as their generalisations (Konishi and Kitagawa, 1996) -, and the evidence, are used in a wide range of contexts, including time series analysis (Shibata, 1976), regression (Hurvich and Tsai, 1989), bias correction (Hurvich and Tsai, 1990), composite likelihoods (Varin and Vidoni, 2005). For a comprehensive treatment of the subject of Model Selection, see e.g. Claeskens and Hjort (2008).

There has been relatively limited research on Model Selection for general classes of HMMs used in practice. A fundamental aspect of a Model Selection Criterion is that of *consistency* (analytically defined later on in the paper). In the HMMs context, Csiszár and Shields (2000) proves strong consistency of BIC for discrete-time, finite-alphabet Markov chains. Gassiat and Boucheron (2003) also considers discrete-time, finite-alphabet HMMs and provides asymptotic and finite-sample analysis of code-based and penalised maximum likelihood estimators (MLEs) using tools from Information Theory and Stein's Lemma. With regards to the Bayesian approach to model selection, this typically involves the marginal likelihood of the data (or evidence) (Jeffreys, 1998; Kass and Raftery, 1995). Shao et al. (2018) show numerically that the evidence can be consistent for HMMs. However, this has yet to be proven analytically.

The work in this paper makes a number of contributions, relevant for HMMs on general state spaces – thus of wide practical significance and such that cover an important gap in the theory of HMMs established in the existing literature. We remark that our analysis assumes smoothness conditions of involved functions w.r.t.˜the parameter of interest, thus is intrinsically not relevant for interesting problems of discrete nature, an example being the identification of the number of states of the underlying Markov chain. Our main results can be summarised as follows:

i) We establish sharp asymptotic results (in the sense of obtaining $\limsup_n$for the quantity of interest) for the log-likelihood function for HMMs evaluated at the MLE, in w.p.1 sense. A lot of the initial developments are borrowed from Douc et al. (2014) (see also citations therein for more work son asymptotic properties of the MLE for HMMs). Moving from the study of the MLE to that of Model Selection Criteria is non-trivial, involving for instance use of the Law of Iterated Logarithm (LIL) for, carefully developed, martingales (Stout, 1970).

ii) We show that BIC and the evidence are strongly consistent in the context of nested HMMs,

whereas AIC is not consistent. To the best of our knowledge, this is the first time that such statements are proven in the literature for general HMMs. For AIC, we show that, w.p. 1, this criterion will occasionally choose the wrong model even under an infinite amount of information.

The rest of the paper is organised as follows. In subsection 5.2, we give basics of some information criteria. Then briefly review some asymptotic results for the log-likelihood function and the MLE without assuming model correctness in subsection 5.3. An important departure from the i.i.d. settings that the log-likelihood function itself does not make up a stationary time-series process even if the data are assumed to be derived from one. subsection 5.4 begins with some asymptotic results for the MLE and the log-likelihood under model correctness. Later on, we move beyond the established literature and, by calling upon LIL for martingales, we establish a number of fundamental asymptotic results, relevant for Model Selection Criteria. In subsection 5.5, we study the derivation of BIC (and its connection with the evidence) and AIC for general HMMs. In particular, an explicit result binding BIC and evidence will later on be used to show that the two criteria share similar consistency properties. subsection 5.6 contains our main results. We prove strong consistency of BIC and the evidence and non-consistency of AIC for a class of nested HMMs. subsection 5.7 reviews (for completeness) an algorithm borrowed from the literature, based on Sequential Monte Carlo, for approximating AIC and BIC. We use this algorithm to present some numerical results that agree with our theory in subsection 5.8. We then conclude in subsection 5.9.

5.2 Basics of information criteria

We first provide a brief explanation of information criteria. Although so many information criteria and their variants have been proposed (see Claeskens and Hjort (2008) for instance), motivated by Gelman et al. (2014), we focus on Akaike information criterion (Akaike, 1974), Bayesian information criterion Schwarz (1978), Deviance information criterion (Spiegelhalter et al., 2002), Watanabe-Akaike information criterion (Watanabe, 2010) and Widely applicable Bayesian information criterion (Watanabe, 2013). Before we proceed to the presentation of these information criteria, we explain the ethos behind them. Roughly speaking, we have chosen to classify the criteria as follows broadly: Frequentist or Bayesian and whether they focus on Kullback–Leibler divergence or model evidence. The first difference comes from the construction of a predictive model. AIC and BIC adopt a frequentist predictive model. In contrast, DIC, WAIC and WBIC use a Bayesian predictive model. We label the first class of information criteria as **Frequentist** and the later ones as **Bayesian**.

The second difference comes from the quantity which an information criterion looks at. Heuristically, let $\hat{p}_\theta(dy_{0:n-1})$ be some predictive model which may or may not depend on the parameter θ. Also let $p_\star(dy_{0:n-1})$ denote the distribution of the data-generating process. Then one might wish to know how close $\hat{p}_\theta(y_{0:n-1})$ is to $p_\star(y_{0:n-1})$ in some sense. We assume that both $\hat{p}_\theta(dy_{0:n-1})$ and $p_\star(dy_{0:n-1})$ admit the densities $\hat{p}_\theta(y_{0:n-1})$, $p_\star(y_{0:n-1})$ w.r.t. dy. In this case, one of natural discrepancy between

$\hat{p}_\theta(y_{0:n-1})$ and $p_\star(y_{0:n-1})$ would be the Kullback–Leibler divergence:

$$\begin{aligned}\mathrm{KL}(\theta) &:= \int p_\star(y_{0:n-1}) \log \frac{p_\star(y_{0:n-1})}{\hat{p}_\theta(y_{0:n-1})} dy_{0:n-1}, \\ &= \int p_\star(y_{0:n-1}) \log p_\star(y_{0:n-1}) dy_{0:n-1} - \int p_\star(y_{0:n-1}) \log \hat{p}_\theta(y_{0:n-1}) dy_{0:n-1}. \end{aligned} \tag{5.1}$$

Critically, the Kullback–Leibler divergence has the desired property such that $\mathrm{KL}(\theta) = 0$ if and only if $\hat{p}_\theta(y_{0:n-1}) = p_\star(y_{0:n-1})$. Therefore $\mathrm{KL}(\theta)$ can been seen as the goodness of fit of a model $\hat{p}_\theta(dy_{0:n-1})$. Notice that minimising (5.1) is equivalent to maximising:

$$\mathcal{R}(\theta) := -\int p_\star(y_{0:n-1}) \log \hat{p}_\theta(y_{0:n-1}) dy_{0:n-1}, \tag{5.2}$$

and hence (5.2) can been seen as a *predictive loss function*. Notice that due to $p_\star(y_{0:n-1})$, one can not calculate (5.2). Therefore if one can construct a proper estimator of (5.2), then this estimator can be understood as an information criterion. As we will present in sequel, AIC, DIC and WAIC are obtained as an estimator of (5.2). We label this class of information criteria as **KL**.

Next, assume that one has two candidate models, say $\mathcal{M}_1$ and $\mathcal{M}_2$. Then Bayesian model comparison can be done via comparing the Bayes factor (Jeffreys, 1998; Kass and Raftery, 1995) between models $\mathcal{M}_1$ and $\mathcal{M}_2$ is given by:

$$BF_{12} := \frac{\int_{\Theta_1} \pi_1(\theta) \exp(\ell_{\theta_1}(y_{0:n-1})) d\theta_1}{\int_{\Theta_2} \pi_2(\theta) \exp(\ell_{\theta_2}(y_{0:n-1})) d\theta_2},$$

where Θ_i and θ_i denote the parameter space and the parameter for the model $i = 1, 2$. From the definition above, it is clear that the key quantity of Bayesian model comparison is the *model evidence* (Jeffreys, 1998):

$$m(y_{0:n-1}) := \int_\Theta \pi(\theta) \exp(\ell_\theta(y_{0:n-1})) d\theta. \tag{5.3}$$

In general, the Bayes factor has the consistency property. That is if one selects the model via the Bayes factor, then, as $n \to \infty$, the selected model is the true one, w.p.1. See Chib and Kuffner (2016) for a general treatment of the consistency of the Bayes factor. However, as one can see, calculating (5.3) involves integrating out θ, and this constrains the application of the model evidence. This is because, in general, such integral cannot be analytically computable and thus (5.3) is computationally expensive, especially for the large d case. Obviously, HMMs are no exception to this problem and would be more problematic due to the expression of the likelihood function in (4.48). To overcome these computational difficulties, BIC and WBIC have been proposed to approximate the log-evidence. We label this class of information criteria as **Evidence**. We summarise our classification in Table 1.

Table 1: Classification of information criteria.

	Frequentist	**Bayesian**
KL	AIC(5.5)	DIC(5.9),WAIC(5.12)
Evidence	BIC(5.7)	WBIC(5.14)

Akaike information criterion (AIC) Akaike information criterion is proposed by Akaike (1974) initially for model selection methods for i.i.d. models and Gaussian models of ARMA type. The essence of AIC is to construct the naive estimator of $\mathcal{R}_n(\theta)$ in (5.2) and correct its bias up to the larger order term (of size $o(1/n)$). Because of asymptotic properties of $\hat{\theta}_{MLE}$, it might be expected that $\mathrm{KL}(\hat{\theta}_{MLE})$ is asymptotically minimised. Thus Akaike (1974) adopts $\frac{1}{n}\ell_{\hat{\theta}_{MLE}}(y_{0:n-1})$ as the the naive estimator of $\mathcal{R}_n(\theta)$. Therefore, AIC can be categorised as a $\mathbf{KL-Frequentist}$ type of information criterion. In order to reduce the bias and appropriately adjust this naive estimator, one would be required an appropriate central limit theorem and thus regular conditions should be satisfied.

Note that AIC does not necessarily require that the parametric model contains the true data distribution. In this case, AIC would turn out to be Takeuchi Information Criterion (TIC), first proposed in Takeuchi (1976). However, in order to derive TIC, the central limit theorem for misspecified models (such as White (1982); Huber (1967)) are typically required.

By making use of asymptotic properties of the MLE, one might obtain the following relation:

$$\mathbb{E}\left[\frac{1}{n}\ell_{\hat{\theta}_{MLE}} - \mathcal{R}_n(\hat{\theta}_{MLE})\right] = \frac{d}{n} + o(n^{-1}), \tag{5.4}$$

where d denotes the number of parameters which the model contain, and this observation implies that the following quantity is the appropriately adjusted naive estimator of $\mathcal{R}_n(\theta)$:

$$AIC := -2\ell_{\hat{\theta}_{MLE}}(y_{0:n-1}) + 2d, \tag{5.5}$$

here we multiplied $\ell_{\hat{\theta}_{MLE}} - d$ by -2 to follow the original definition in Akaike (1974). Therefore, the model which has the minimum AIC value can be considered as the best model among candidate models.

As we shall explain later, it is well known that AIC tends to select an over-fitting model. That is, AIC does not necessarily select the true model. To overcome this problem, some author proposes high-order bias correction methods. However, these modified types of AIC are rarely used in practice. In the context of HMMs, see Bengtsson and Cavanaugh (2006) for instance. We do not cover this direction in this study.

Bayesian information criterion (BIC) In general, evaluating evidence involves computational techniques since one has to do integrating out parameters from the joint density of $(\theta, y_{0:n-1})$. This is also the case for HMMs. See, for instance, Zhou et al. (2016) for making use of Sequential Monte Carlo samplers, Chib (1995) for MCMC approach and Gelman and Meng (1998) for via importance, bridge and path sampling methods. In contrast with such computational methods, Schwarz (1978) makes directly use of the Laplace approximation to $\log m(y_{0:n-1})$ and this leads Bayesian information criterion (BIC).

Note that using Laplace approximation requires some additional conditions on the likelihood and the prior density and the models satisfying such conditions be often called Laplace-regular models, see Kass et al. (1990) for details. Roughly speaking, Laplace-regular requires (1) high-order continuity and differentiability of the log-likelihood function and prior density, (2) uniform bound for derivatives

of the log-likelihood function, (3) uniform convergence of the log-likelihood function and (4) uniform convergence of the observed Fisher information matrix. After some careful calculations, one will end up with the following expression:

$$\frac{m(y_{0:n-1})}{p_{\hat{\theta}_{MLE}}(y_{0:n-1})} = (2\pi)^{d/2} n^{d/2} \left[\det\left(-\frac{1}{n}\nabla_\theta \nabla_\theta^\top \ell_{\hat{\theta}_{MLE}}(y_{0:n-1})\right)\right]^{-1/2} \pi(\hat{\theta}_{MLE})(1+\mathcal{O}(n^{-1})). \tag{5.6}$$

Under the assumptions, one can show that $\left[\det\left(-\frac{1}{n}\nabla_\theta \nabla_\theta \ell_{\hat{\theta}_n}(y_{0:n-1})\right)\right]^{-1/2}$, and $\pi(\hat{\theta}_{MLE})$ are $\mathcal{O}(1)$. Ignoring these $\mathcal{O}(1)$ terms implies that $\log m(y_{0:n-1})$ can be approximated by $\ell_{\hat{\theta}_{MLE}}(y_{0:n-1}) - \frac{d\log n}{2}$ and this gives rise to BIC:

$$BIC := -2\ell_{\hat{\theta}_{MLE}}(y_{0:n-1}) + d\log n, \tag{5.7}$$

here we multiplied $\ell_{\hat{\theta}_{MLE}}(y_{0:n-1}) - \frac{d\log n}{2}$ by -2 so as to be comparable with AIC. From above presentation and definition of BIC, it is obvious that BIC can be categorised as a **Evidence** − **Frequentist** type of information criterion. Also, it is clear that BIC is not **Bayesian** type of information criterion, and has a bit misleading name (Gelman et al., 2014), even though 'B' in BIC is for Bayesian.

Although the only difference between BIC and AIC is the penalty term, that is, $d\log n$ for BIC and $2d$ for AIC, in contrast with AIC, it is also well known that the BIC is strongly consistent in i.i.d. settings and some particular non-i.i.d. ones, e.g. Claeskens and Hjort (2008); Nishii (1988).

Deviance information criterion (DIC) As we will observe later, Deviance information criterion (DIC), first studied in Spiegelhalter et al. (2002), may be understood as a Bayesian version of AIC in (5.5).

In contrast with Akaike (1974), Spiegelhalter et al. (2002) work on the log-likelihood evaluated at the posterior mean in (4.52) (the posterior mode in (4.53) can be also available), that is $\ell_{\bar{\theta}_n}(y_{0:n-1})$. Then they define the *deviance* $\mathcal{D}(y_{0:n-1},\theta) := -2\ell_\theta(y_{0:n-1})$ and consider the difference of the deviance between the posterior mean and the parameter, that is:

$$\begin{aligned}\mathcal{DD}(y_{0:n-1},\theta,\bar{\theta}_n) &:= \mathcal{D}(y_{0:n-1},\theta) - \mathcal{D}(y_{0:n-1},\bar{\theta}_n),\\ &= 2\left(-\ell_\theta(y_{0:n-1}) + \ell_{\bar{\theta}_n}(y_{0:n-1})\right),\end{aligned}$$

here θ is a random variable drawing from a prior $\pi(\theta)$. To evaluate complexity of a model, Spiegelhalter et al. (2002) propose the following measure as the effective number of parameters:

$$\begin{aligned}p_{DIC} &:= \mathbb{E}_\Pi\left[\mathcal{DD}(y_{0:n-1},\theta,\bar{\theta}_n)\right],\\ &= 2\ell_{\bar{\theta}_n}(y_{0:n-1}) - 2\int_\Theta \ell_\theta(y_{0:n-1})\Pi(\theta \mid y_{0:n-1})d\theta,\end{aligned} \tag{5.8}$$

and then DIC is defined by:

$$DIC := \mathcal{D}(y_{0:n-1},\hat{\theta}_{PM}) + 2p_{DIC}.$$

Note that $-2\int_\Theta \ell_\theta(y_{0:n-1})\Pi(\theta \mid y_{0:n-1})d\theta + p_{DIC} = 2\ell_{\bar{\theta}_n}(y_{0:n-1}) - 4\int_\Theta \ell_\theta(y_{0:n-1})\Pi(\theta \mid y_{0:n-1})d\theta$. Thus, DIC can alternatively be defined as:

$$DIC := -2\int_\Theta \ell_\theta(y_{0:n-1})\Pi(\theta \mid y_{0:n-1})d\theta + p_{DIC}. \tag{5.9}$$

Derivation of DIC in Spiegelhalter et al. (2002) is somewhat heuristic, however, as Ando (2007) points out, it can been seen that Spiegelhalter et al. (2002) implicitly obtain DIC as the maximisation of the posterior mean of the expected log-likelihood, that is $\mathbb{E}_\Pi[\mathcal{R}_n(\bar{\theta}_n)]$. Therefore, DIC would be understood as a Bayesian variant of AIC and thus, categorised as a $\mathbf{KL-Bayesian}$ type of information criterion.

Although DIC seems to lack of the theoretical meaning, it has been widely used in Bayesian data analysis due to its flexibility and computational efficiency. As for HMMs, Berg et al. (2004) use DIC for comparing the performance of a variety of stochastic volatility models and Yu and Meyer (2006) compare multivariate stochastic volatility models using DIC. Critically, same as AIC, DIC is known to be lack of the consistency property. Although Spiegelhalter et al. (2002) provides only heuristic analysis, it can be seen that DIC has the same asymptotic behaviour as AIC as a consequence of the Bernstein–von Mises type theorem. Hence, DIC and AIC would be in agreement, and this leads to the deficiency of consistency property. In the case of HMMs, this asymptotic equivalence has yet to been proven. Another problem concerning DIC is instability of p_{DIC} in (5.8) in the sense that DIC in (5.9) ends up with a negative value of p_{DIC} . As pointed out in Brooks et al. (2002), this shortcoming would remarkably appear in the model with latent variables such as mixture models and, of course, HMMs. From a theoretical point of view, this phenomenon comes from the identifiability problem and this task appears to be notoriously challenging in the case of HMMs, see Allman et al. (2009); Douc et al. (2014) and references therein.

Watanabe-Akaike information criterion (WAIC) Recall that deriving AIC (also BIC) involves the asymptotic normality of the MLE. In general, such asymptotic result requires (1) convergence of the score function and (2) convergence of the observed Fisher information. Then the later convergence result typically accompanies the assumption such that the observed Fisher information is non-singular. In HMMs context, although the rigorous proof and the derivation are rather technical, this type of the limit theorem has been established and well studied (Douc et al., 2004, 2014).

The key contribution of WAIC, presented in Watanabe (2010), is that one does not need to assume that the observed Fisher information is non-singular. That is, WAIC works even for singular models. This is the desirable property of WAIC since checking non-singularity of the observed Fisher information is a somewhat a priori assumption, and it might not be possible to do in advance, in general. Also as Watanabe (2010) pointed out, such singularity might arise potentially in latent variable models such as mixture models, neural networks, hierarchical models and, critically, HMMs.

Critically, Watanabe (2010) argues that if the observed Fisher information is singular, then AIC is not asymptotically an unbiased estimator of (5.2) in the i.i.d. setting. Instead of the plug-in predictive

density, WAIC first considers:

$$\hat{p}(y_n \mid y_{0:n-1}) := \int p_\theta(y_n \mid y_{0:n-1})\Pi(\theta \mid y_{0:n-1})d\theta \tag{5.10}$$

as the point-wise Bayesian predictive density. Also, instead of (5.2), WAIC works with:

$$G_n(\theta) := -\int p_\star(y_{0:n-1}) \log \hat{p}(y_{0:n-1})dy_{0:n-1}, \tag{5.11}$$

as a loss function, here $\hat{p}(y_{0:n-1}) := \prod_{k=1}^{n-1} \hat{p}(y_k \mid y_{0:k-1})$. That is, Watanabe (2010) considers the Kullback–Leibler divergence between $p_\star(y_{0:n-1})$ and $\hat{p}(y_{0:n-1})$, and hence (5.11) can been seen as a predictive loss function from a fully Bayesian view. Then, by making use of some techniques from algebraic geometry, the following might be an asymptotically unbiased estimator of (5.11), even if models are singular:

$$WAIC := -2\sum_{t=0}^{n-1} \log \mathbb{E}_\theta \left[p_\theta(y_t \mid y_{0:t-1}) \mid y_{0:n-1}\right] + 2\sum_{t=0}^{n-1} \mathbb{V}_\theta[\log p_\theta(y_t \mid y_{0:t-1}) \mid y_{0:n-1}]. \tag{5.12}$$

We note that our definition of WAIC in (5.12) is different from the one in Watanabe (2010). In Watanabe (2010), the definition of WAIC is given by:

$$-2\sum_{t=0}^{n-1} \log \mathbb{E}\left[p_\theta(y_t) \mid y_{0:n-1}\right] + 2\sum_{t=1}^{n-1} \mathbb{V}\left[\log p_\theta(y_t) \mid y_{0:n-1}\right].$$

That is, WAIC partitions the data into n pieces and this is one of the critical difficulties to make use of WAIC (Gelman et al., 2014). In particular, one cannot define WAIC in such way for HMMs because of dependency of data. Hence, we have heuristically defined WAIC for HMMs as in (5.12). From discussion above, one can see that WAIC can be understood as a fully Bayesian generalisation of AIC, hence can be classed as a **KL − Bayesian** type of information criterion. Note that, same as AIC, deriving WAIC for HMMs itself would be rather challenging. Our presentation here is to facilitate study and heuristic. Rigorous derivation has to be done. Interestingly, Watanabe (2010) points out that if the models being considered are not singular one, then the averages of WAIC, AIC and DIC have the same asymptotic behaviour, but such result has yet to been proven in the context of HMMs. Note that here we have defined WAIC as $-2n$ times the original definition in Watanabe (2010) so as to have the same scale in AIC and DIC.

Widely applicable Bayesian information criterion (WBIC) The presentation in this section follows closely Friel et al. (2017). WBIC (Watanabe, 2013) shares the same motivation as in Watanabe (2010). Recall that BIC is obtained as the Laplace approximation of log-evidence, and, critically, the Laplace approximation needs models that are Laplace-regular. If these conditions fail, log-evidence may not be asymptotically approximated by BIC and such approximation is known to be poor (Chickering and Heckerman, 1997). That is, BIC may not be a decent approximation of log-evidence for singular models, and WBIC is designed to address this issue. As the sample size $n \to \infty$, Watanabe (2013) shows that, in the i.i.d. setting, WBIC converges to the log-evidence even if the models being considered

are singular.

Let $\{\phi_n\}$ be a sequence of the inverse temperatures such that $0 < \phi_n < \phi_{n-1} < \cdots \phi_0 = 1$. Then, as in annealed importance sampling (Neal, 2001), Watanabe (2013) considers the annealed posterior density such that:

$$\Pi(\theta \mid y_{0:n-1})^{\phi_n} := \frac{p_\theta(y_{0:n-1})^{\phi_n}\pi(\theta).}{\int_\Theta p_\theta(y_{0:n-1})^{\phi_n}\pi(\theta)d\theta}. \tag{5.13}$$

Clearly, $\Pi(\theta \mid y_{0:n-1})^{\phi_0} = \Pi(\theta \mid y_{0:n-1})$ holds. Then one can show that there exists the optimal schedule $\{\phi_n^*\}$ in the sense that $\log m(y_{0:n-1}) = \mathbb{E}_{\Pi^{\phi_n^*}}[\ell_\theta(y_{0:n-1})]$ holds for any n, where $\mathbb{E}_{\Pi^{\phi_n^*}}[\cdot]$ denotes the expectation w.r.t. the the annealed posterior density in (5.13). Although identifying this optimal schedule $\{\phi_n^*\}$ is a challenging task, Watanabe (2013) shows that the choice $\phi_n = 1/\log n$ is asymptotically equivalent to ϕ_n^* for singular models. Hence this observation leads to WBIC:

$$WBIC := \frac{\int_\Theta \ell_\theta(y_{0:n-1})p_\theta(y_{0:n-1})^{1/\log n}\pi(\theta)d\theta}{\int_\Theta p_\theta(y_{0:n-1})^{1/\log n}\pi(\theta)d\theta}. \tag{5.14}$$

In other words, WBIC is defined as the annealed posterior mean of the log-likelihood with the choice $\phi_n = 1/\log n$. First notice that, given ϕ_n^{WBIC}, one can estimate $\log m(y_{0:n-1})$ via one simulation done by, for instance, MCMC. Hence, WBIC successfully reduces computational cost. Although Watanabe (2013) does not show it, WBIC might have the consistent property since it might be shown that WBIC and BIC are asymptotically equivalent for non-singular models. Also, as pointed by Friel et al. (2017), there has been limited (numerical) exploration of this criterion, see Mononen (2015) for a Gaussian process regression, Friel et al. (2017) for some simple tractable models. In Watanabe (2013), it is numerically presented that WBIC works better than BIC for a reduced rank regression model. Again, our presentation here is heuristic. Although we have not tried to derive WBIC for HMMs, as this would make this part of the text appear overly cumbersome, the rigorous derivation of WBIC for HMMs is itself rather important. From the discussion above, we categorise WBIC as a **Evidence** – **Bayesian** type of information criterion.

Remark 8. *Critically,* **KL** *and* **Evidence** *apparently have different goals. As we presented,* **KL** *tries to estimate* (5.1) *whereas* **Evidence** *tries to estimate the (log) evidence in* (5.3)*. To be precise,* **KL** *type criteria evaluate models from the viewpoint of prediction and* **Evidence** *type criteria focus on the posterior probabilities of models. Therefore, theoretically speaking, we recognise that comparing* **KL** *type criteria and* **Evidence** *type ones is somewhat misleading. However, in practice, we think that it would be desirable that one compare the values of these different criteria and selects a model based on them if it is possible.*

5.3 Asymptotics under no-model-correctness

We briefly summarise some asymptotic results for general HMMs needed in later sections. The development follows closely Douc et al. (2014, Chapter 13) Again, an HMM is a bivariate process $\{x_k, z_k\}_{k\geq 0}$ such that state component $\{x_k\}_{k\geq 0}$ is an unobservable Markov chain with initial law $x_0 \sim \eta$ and transition kernel $Q_\theta(\cdot \mid x)$, with values in the measurable space $(\mathsf{X}, \mathcal{X})$. We have adopted a parametric

setting with $\theta \in \Theta \subseteq \mathbb{R}^d$, for some $d \geq 1$. Conditionally on $\{x_k\}_{k\geq 0}$, the distribution of the observation process instance $z_k = z$ depends only on $x_k = x$, independently over $k \geq 0$, and is given by the kernel $G_\theta(\cdot \mid x)$defined on $(\mathsf{Y}, \mathcal{Y})$. We assume that X and Y are Polish spaces and $\mathcal{X}$, $\mathcal{Y}$ the corresponding Boreal-algebras. The notation$\{y_k\}_{k\geq 0}$ is reserved for the true data generating process, which may or may not belong in the parametric family of HMMs we specified above – meant to be distinguished from $\{z_k\}_{k\geq 0}$ which is the process driven by the model dynamics. In particular, in this section we work under no-model correctness, i.e. we do not have to assume the existence of a correct parameter value for the prescribed model that delivers the distribution of the data generating process.

Throughout the article, we assume that the following hold.

Assumption 1. *The data generating process $\{y_k\}_{k\geq 0}$ is strictly stationary and ergodic.*

Assumption 2. *i) There exists a probability measure μ on $(\mathsf{X}, \mathcal{X})$ which dominates the kernel $Q_\theta(\cdot \mid x)$ for any $(x, \theta) \in \mathsf{X} \times \Theta$ with density $q_\theta(x' \mid x) := (dQ_\theta(\cdot \mid x)/d\mu)(x')$.*

ii) There exists a probability measure ν on $(\mathsf{Y}, \mathcal{Y})$ which dominates the kernel $G_\theta(\cdot \mid x)$ for any $(x, \theta) \in \mathsf{X} \times \Theta$ with density $g_\theta(y \mid x) := (dG_\theta(\cdot \mid x)/d\nu)(y)$.

iii) The parameter space Θ is a compact subset of $\mathbb{R}^d$; w.p.1, $p_\theta(y_{0:n-1}) > 0$ for all $\theta \in \Theta$, for all n, where $p_\theta(\cdot)$ denotes here the density of the distribution of the observations under the model (for given θ and size n).

iv) The initial distribution $\eta(dx_0)$ has the density, denoted $\eta(x_0)$, w.r.t. μ; it also has finite first moment.

Without loss of generality, we have assumed that $\eta(dx_0)$ does not depend on θ. Probability statements -as in Assumption 2 (*iii*)- and expectations throughout the paper are to be understood w.r.t. the law of the data generating process $\{y_k\}$. Henceforth we make use of the notation $a_{i:j} = (a_i, \cdots, a_j)$, for integers $i \leq j$, for a given sequence $\{a_k\}$. We need the following technical assumptions.

Assumption 3. *There exists $\sigma^-, \sigma^+ \in (0, \infty)$ such that*

$$\sigma^- \leq q_\theta(x' \mid x) \leq \sigma^+$$

for any $x, x' \in \mathsf{X}$ and any $\theta \in \Theta$.

Assumption 3 is the 'strong mixing condition' typically used in this context (Cappé et al., 2005; Del Moral, 2004), providing a Dobrushin coefficient of $1 - \frac{\sigma^-}{\sigma^+}$ for the hidden Markov chain; it is critical for most of the results reviewed or developed in the sequel. Assumption 3 implies, for instance, that for any $x \in \mathsf{X}$, $A \in \mathsf{X}$, $Q_\theta(A \mid x) \geq \sigma^- \mu(A)$, that is, for any $\theta \in \Theta$, X is a 1-small set for process $\{x_k\}_{k\geq 0}$. Thus, the chain has the unique invariant measure π_θ^X and is uniformly ergodic, so for any $x \in \mathsf{X}$, $n \geq 0$, $\left\|Q_\theta^n(\cdot \mid x) - \pi_\theta^X\right\|_{TV} \leq \left(1 - \frac{\sigma^-}{\sigma^+}\right)^n$ with $\|\cdot\|_{TV}$ denoting the total variation norm.

We calculate the likelihood and log-likelihood functions:

$$p_\theta(y_{0:n-1}) := \int \eta_\theta(x_0) g_\theta(y_0 \mid x_0) \prod_{k=1}^{n-1} g_\theta(y_k \mid x_k) q_\theta(x_k \mid x_{k-1}) \mu^{\otimes n}(dx_{0:n-1}), \tag{5.15}$$

$$\ell_\theta(y_{0:n-1}) := \log p_\theta(y_{0:n-1}) = \sum_{k=0}^{n-1} \log p_\theta(y_k \mid y_{0:k-1}). \tag{5.16}$$

Though $\{y_k\}_{k\geq 0}$ is stationary and ergodic, the terms $\{\log p_\theta(y_k \mid y_{0:k-1})\}_{k\geq 0}$ do not form a strictly stationary process (in general). To obtain (strictly) stationary and ergodic log-likelihood terms, following Douc et al. (2004, 2014); Cappé et al. (2005), we work with the standard extension of the $y-$process onto the whole of integers, and write $\{y_k\}_{k=-\infty}^{\infty}$. One can then define the variable $\log p_\theta(y_k \mid y_{-\infty:k-1})$ as the $w.p.1$ limit of the Cauchy sequence (uniformly in θ) $\log p_\theta(y_k \mid y_{-t:k-1})$, found as in (5.15) for initial law $x_{-t} \sim \eta$ as $t \to \infty$; see Douc et al. (2014, Chapter 13) for more details. We can now define the modified, stationary version of the log-likelihood:

$$\ell_\theta^s(y_{-\infty:n-1}) := \sum_{k=0}^{n-1} \log p_\theta(y_k \mid y_{-\infty:k-1}), \tag{5.17}$$

where superscript s stands for stationary. We will need the following assumption.

Assumption 4. *We have that* $b^+ := \sup_\theta \sup_{x,y} g_\theta(y \mid x) < \infty$ *and:*

$$\mathbb{E}\left[\left|\log b^-(y_0)\right|\right] < \infty,$$

where $b^-(y) := \inf_\theta \int_{\mathsf{X}} g_\theta(y \mid x)\mu(dx)$.

The finite-moment part implies that $\mathbb{E}\left[\log p_\theta(y_0 \mid y_{-\infty:-1})\right] < \infty$, thus Birkhoff's ergodic theorem can be applied for averages deduced from (5.17).

Proposition 26. *Under Assumption 1 to Assumption 4,*

$$\sup_{\theta\in\Theta} \left| \frac{1}{n}\ell_\theta(y_{0:n-1}) - \frac{1}{n}\ell_\theta^s(y_{-\infty:n-1}) \right| \leq \frac{C}{n},$$

for a constant $C > 0$.

Proof. This is Proposition 13.5 of Douc et al. (2014); the upper bound C/n is implied from the proof of that proposition. □

We consider the maximum likelihood estimator defined as the set:

$$\hat{\theta}_n := \arg\max_{\theta\in\Theta} \ell_\theta(y_{0:n-1}). \tag{5.18}$$

We make the following assumption.

Assumption 5. *For all* $(x, x') \in \mathsf{X}\times\mathsf{X}$ *and* $y \in \mathsf{Y}$, *the mappings* $\theta \mapsto q_\theta(x' \mid x)$ *and* $\theta \mapsto g_\theta(y \mid x)$ *are continuous.*

Such conditions imply continuity for the log-likelihood mapping $\theta \mapsto \frac{1}{n}\ell_\theta(y_{0:n-1})$ and its limit $\theta \mapsto \mathbb{E}\left[\log p_\theta(y_0 \mid y_{-\infty:-1})\right]$, which, together with with other conditions, provide convergence of the MLE to the maximiser of the limiting function. For sets $A, B \subseteq \Theta$, we define $d(A, B) := \inf_{a\in A,\, b\in B} |a - b|$.

Proposition 27. *Under Assumption 1 to Assumption 5, we have the followings.*

i) Let $\ell(\theta) := \mathbb{E}\left[\log p_\theta(y_0 \mid y_{-\infty:-1})\right]$. *The function* $\theta \mapsto \ell(\theta)$ *is continuous, and we have that:*

$$\lim_{n\to\infty} \sup_{\theta\in\Theta} \left| \frac{1}{n}\ell_\theta(y_{0:n-1}) - \ell(\theta) \right| = 0, \ \ w.p.1.$$

ii) We have $\lim_{n\to\infty} d(\hat{\theta}_n, \theta_\star)$, *w.p.*1, *where*

$$\theta_\star := \arg\max_{\theta\in\Theta} \ell(\theta)$$

is the set of global maxima of $\ell(\theta)$.

Proof. This follows from Proposition 13.7 of Douc et al. (2014). The proof of the first statement is based on working with the stationary version of the log-likelihood in (5.17), permitted due to Proposition 27, and using Birkhoff's ergodic theorem. □

One can see Proposition 27 as a generalisation of the Shannon-McMillan-Breiman theorem. Recall that $\theta_\star$ need not be thought of as the correct parameter value here, as no assumption of the class of HMMs containing the correct data-generating model is made in this section. To avoid identifiability issues, we make the following assumption on the HMM model. We refer the reader to Theorem 13.14 in Douc et al. (2014) and discussions therein.

Assumption 6. $\theta_\star$ *is a singleton.*

This implies immediately the following.

Corollary 1. *The set of maxima* $\hat{\theta}_n$ *is a singleton for all large enough* n. *Therefore, we obtain* $\lim_{n\to\infty} \hat{\theta}_n = \theta_\star$ *w.p.1.*

5.4 Asymptotics under model-correctness

To examine the asymptotic behaviour of the AIC or BIC, one has to investigate the behaviour of the log-likelihood evaluated at the MLE, $\ell_{\hat{\theta}_n}(y_{0:n-1})$, for increasing sample size n. Following closely Douc et al. (2004, 2014), we first pose the following assumption, with $\theta_\star \in \mathbb{R}^d$ as determined in Proposition 27 and Assumption 6. Here and in the sequel, all gradients and Hessians - represented by ∇ and $\nabla\nabla^\top$ respectively adopting an 'applied mathematics' notation - are w.r.t. the model parameter(s) θ.

Assumption 7. θ_* *is in the interior of* Θ, *and exist* $\epsilon > 0$ *and an open neighbourhood* $\mathcal{B}_\epsilon(\theta_\star) := \{\theta \in \Theta :| \theta - \theta_\star |< \epsilon\}$ *of* $\theta_\star$ *such that the following hold.*

i) For any (x, x') *in* $\mathsf{X} \times \mathsf{X}$ *and* $y \in \mathsf{Y}$, $\theta \mapsto q_\theta(x' \mid x)$ *and* $\theta \mapsto g_\theta(y \mid x)$ *are twice differentiable on* $\mathcal{B}_\epsilon(\theta_\star)$.

ii) $\sup_{\theta\in\mathcal{B}_\epsilon(\theta_\star)}\sup_{x,x'\in\mathsf{X}^2}\left\{\|\nabla\log q_\theta(x'\mid x)\|+\left\|\nabla\nabla^\top\log q_\theta(x'\mid x)\right\|\right\}<\infty$.

iii) For some $\delta>0$,

$$\mathbb{E}\left[\sup_{\theta\in\mathcal{B}_\epsilon(\theta_\star)}\sup_{x\in\mathsf{X}}\left\{\|\nabla\log g_\theta(y_0\mid x)\|_{2+\delta}+\nabla\nabla^\top\log g_\theta(y_0\mid x)\right\}\right]<\infty.$$

$|\cdot|$ denotes the Euclidean norm for vector input or one of the standard equivalent matrix norms for matrix input. Assumption 7 can be seen as a natural extension of regular conditions to prove asymptotic normality of the MLE in the case of HMMs. That is, for any fixed n, the log-likelihood function is twice continuously differentiable in $\mathcal{B}_\epsilon(\theta_\star)$ (standard use of dominated convergence theorem from (1) of Assumption 7). Also, the score function has finite $(2+\delta)-$moment and the Hessian finite first moment, for any $\theta\in\mathcal{B}_\epsilon(\theta_\star)$; the proof requires use of Fisher's identity (used later on) and parts (2), (3) of Assumption 7 involving the gradient for the score function and Louis' identity for the Hessian together with the stated conditions for the matrices of second derivatives. We avoid further details.

We start off with a standard Taylor expansion (see Appendix A):

$$\begin{aligned}\ell_{\hat\theta_n}(y_{0:n-1})=\ell_{\theta_\star}(y_{0:n-1})+\frac{1}{\sqrt{n}}\nabla\ell_{\theta_\star}(y_{0:n-1})\sqrt{n}\left(\hat\theta_n-\theta_\star\right)\\ +\frac{1}{2}\sqrt{n}\left(\hat\theta_n-\theta_\star\right)^\top\left[\int_0^1\frac{\nabla\nabla^\top\ell_{s\hat\theta_n+(1-s)\theta_\star}(y_{0:n-1})ds}{n}s\right]\sqrt{n}\left(\hat\theta_n-\theta_\star\right),\end{aligned}\tag{5.19}$$

together with a corresponding one for the score function,

$$\begin{aligned}0\equiv\frac{1}{\sqrt{n}}\nabla\ell_{\hat\theta_n}(y_{0:n-1})=\frac{1}{\sqrt{n}}\nabla\ell_{\theta_\star}(y_{0:n-1})\\ +\left[\int_0^1\frac{\nabla\nabla^\top\ell_{s\hat\theta_n+(1-s)\theta_\star}(y_{0:n-1})ds}{n}\right]\sqrt{n}\left(\hat\theta_n-\theta_\star\right).\end{aligned}\tag{5.20}$$

We will look at the asymptotic properties of the score function terms and the integral involving the Hessian, i.e. of,

$$\frac{1}{\sqrt{n}}\ell_{\theta_\star}(y_{0:n-1}),\int_0^1\frac{1}{n}\nabla\nabla^\top\ell_{s\hat\theta_n+(1-s)\theta_\star}(y_{0:n-1})ds\tag{5.21}$$

starting from the former.

We will sometimes need to work under the assumption of model-correctness and we shall be clear when that is the case.

Assumption 8. *The dynamics of the data generating process* $\{y_k\}_{k\geq0}$ *corresponds to those of the HMM with initial distribution* $x_0\sim\eta(\cdot)=\pi^X_{\theta_\star}(\cdot)$*, transition kernel* $Q_{\theta_\star}(\cdot\mid x)$ *and kernel* $G_{\theta_\star}(\cdot\mid x)$.

For results that do no refer to Assumption 8, still makes sense as per its definition in Proposition 27. Using Jensen's inequality, and for $\theta_\star$ corresponding to the true parameter, one can easily check that $\ell(\theta)\leq\ell(\theta_\star)$, so indeed the true parameter coincides with $\theta_\star$ given in Proposition 26.

Following Douc et al. (2014, Chapter 13), we proceed with the following 4 steps.

Step 1. Re-write the score function evaluated at $\theta = \theta_\star$ as:

$$\nabla \ell_{\theta_\star}(y_{0:n-1}) = \sum_{i=0}^{n-1} \left[\nabla \ell_{\theta_\star}(y_{0:i}) - \nabla \ell_{\theta_\star}(y_{0:i-1}) \right], \tag{5.22}$$

under the convention $\nabla \ell_{\theta_\star}(y_{0:-1}) \equiv 0$. The above differences will be shown to converge - for increasing sample size n, in an appropriate sense - to stationary (and ergodic) martingale increments.

Step 2. Using the Fisher's identity, one has, for $y_{0:k} \in \mathsf{Y}^{k+1}$, $k \geq 0$:

$$\begin{aligned} \nabla \ell_{\theta_\star}(y_{0:k}) &= \int_{\mathsf{X}^{k+1}} \nabla p_{\theta_\star}\left(x_{0:k}, y_{0:k}\right) p_{\theta_\star}\left(x_{0:k} \mid y_{0:k}\right) \mu^{\otimes(k+1)}(dx_{0:k}), \\ &= \sum_{j=0}^{k} \int_{\mathsf{X}^2} d_{\theta_\star}\left(x_{j-1}, x_j, y_j\right) p_{\theta_\star}\left(x_{j-1:j} \mid y_{0:k}\right) \mu^{\otimes 2}(dx_{j-1:j}),\ j \geq 0, \end{aligned} \tag{5.23}$$

where we have defined as:

$$d_{\theta_\star}\left(x_{j-1}, x_j, y_j\right) := \nabla \log \left[q_{\theta_\star}(x_j \mid x_{j-1}) g_{\theta_\star}(y_j \mid x_j) \right],$$

with the conventions:

$$d_{\theta_\star}(x_{-1}, x_0, y_0) \equiv d_{\theta_\star}(x_0, y_0) \equiv \nabla \log \left[\eta(x_0) g_{\theta_\star}\left(y_0 \mid x_0\right) \right],$$

and the one:

$$\begin{aligned} &\int_{\mathsf{X}^2} d_{\theta_\star}(x_{-1}, x_0, y_0) p_{\theta_\star}\left(x_{-1:0} \mid y_{0:k}\right) \mu^{\otimes 2}(dx_{-1:0}) \\ &\qquad \equiv \int_{\mathsf{X}^2} d_{\theta_\star}(x_0, y_0) p_{\theta_*}\left(x_0 \mid y_{0:k}\right) \mu(dx_0). \end{aligned}$$

Thus, we have for $i \geq 0$,

$$\begin{aligned} h_{\theta_\star}(y_{0:i}) &:= \nabla \ell_{\theta_\star}(y_{0:i}) - \nabla \ell_{\theta_\star}(y_{0:i-1}), \\ &= \int_{\mathsf{x}^2} d_{\theta_\star}\left(x_{i-1,} x_i, y_i\right) p_{\theta_\star}\left(x_{i-1:i} \mid y_{0:i}\right) \mu^{\otimes 2}(dx_{i-1:i}) \\ &+ \sum_{j=0}^{i-1} \left[\int_{\mathsf{x}^2} d_{\theta_\star}\left(x_{j-1}, x_j, y_j\right) p_{\theta_\star}\left(x_{j-1:j} \mid y_{0:i}\right) \mu^{\otimes 2}(dx_{j-1:j}) \right. \\ &\left. \int_{\mathsf{x}^2} d_{\theta_\star}\left(x_{j-1}, x_j, y_j\right) p_{\theta_\star}\left(x_{j-1:j} \mid y_{0:i-1}\right) \mu^{\otimes 2}(dx_{j-1:j}) \right]. \end{aligned} \tag{5.24}$$

Note that since we have assumed that $\{y_k\}_{k \geq 0}$ is strictly stationary and ergodic, this leads one to

extend, for any integers $i \geq 1$ and $m \geq 0$, $h_{\theta_\star}(y_{0:i})$ to:

$$h_{\theta_\star}(y_{-m:i}) = \int_{\mathsf{x}^2} d_{\theta_\star}(x_{i-1}, x_i, y_i)\, p_{\theta_\star}(x_{i-1:i} y_{-m:i})\, \mu^{\otimes 2}(dx_{i-1:i}) + \sum_{j=-m+1}^{i-1} \left[\int_{\mathsf{x}^2} d_{\theta_\star}(x_{j-1}, x_j, y_j)\, p_{\theta_\star}(x_{j-1:j} \mid y_{-m:i})\, \mu^{\otimes 2}(dx_{j-1:j}) \int_{\mathsf{x}^2} d_{\theta_\star}(x_{j-1}, x_j, y_j)\, p_{\theta_\star}(x_{j-1:j} \mid y_{-m:i-1})\, \mu^{\otimes 2}(dx_{j-1:j}) \right]. \tag{5.25}$$

Following Douc et al. (2014, Proposition 13.20), integrals involving infinitely long data sequences of the form $\int_{\mathsf{x}^2} d_{\theta_\star}(x_{j-1}, x_j, y_j)\, p_{\theta_\star}(x_{j-1:j} \mid y_{-\infty:i})\, \mu^{\otimes 2}(dx_{j-1:j})$, $j \leq i$, $i > 0$ can be defined as w.p.1or L^2-limits of the random variables $\int_{\mathsf{x}^2} d_{\theta_\star}(x_{j-1}, x_j, y_j)\, p_{\theta_\star}(x_{j-1:j} \mid y_{-m:i})\, \mu^{\otimes 2}(dx_{j-1:j})$ with $m \to \infty$, under Assumption 1 to Assumption 7.

Step 3. Then letting $m \to \infty$, we have the limit as:

$$\int_{\mathsf{x}^2} d_{\theta_\star}(x_{j-1}, x_j, y_j)\, p_{\theta_\star}(x_{j-1:j} \mid y_{-\infty:i})\, \mu^{\otimes 2}(dx_{j-1:j}).$$

Thus now we can define the limit of $h_{\theta_\star}(y_{-\infty:i})$ as follows:

$$h_{\theta_*}(y_{-\infty:i}) := \int_{\mathsf{x}^2} d_{\theta_\star}(x_{i-1}, x_i, y_i)\, p_{\theta_\star}(x_{i-1:i} \mid y_{-\infty:i})\, \mu^{\otimes 2}(dx_{i-1:i}) + \sum_{j=-\infty+1}^{i-1} \left[\int_{\mathsf{x}^2} d_{\theta_\star}(x_{j-1}, x_j, y_j)\, p_{\theta_\star}(x_{j-1:j} \mid y_{-\infty:i})\, \mu^{\otimes 2}(dx_{j-1:j}) \int_{\mathsf{x}^2} d_{\theta_\star}(x_{j-1}, x_j, y_j)\, p_{\theta_\star}(x_{j-1:j} \mid y_{-\infty:i-1})\, \mu^{\otimes 2}(dx_{j-1:j}) \right]. \tag{5.26}$$

A small modification of the derivations in Douc et al. (2014, Lemma 13.21), (they look at the second moment) gives that, under Assumption 1-Assumption 7 and for the constant $\delta > 0$ as defined in (3) of Assumption 7, for $i \geq 0$,

$$\|h_{\theta_\star}(y_{0:i}) - h_{\theta_*}(y_{-\infty:i})\|_{2+\delta} \leq 12 \left(\mathbb{E}\left[\sup_{x,\, x' \in \mathsf{X}} d_{\theta_\star}(x, x', y) \right]^{2+\delta} \right)^{\frac{1}{2+\delta}} \frac{\rho^{i/2-1}}{1-\rho}, \tag{5.27}$$

where we have defined $\rho := 1 - \frac{\sigma^-}{\sigma^+}$. (The expectation in the upper bound is finite due to (2), (3) of Assumption 7). Let quantity $\|\cdot\|_a$, $a \geq 1$ denote the L^a norm of the variable under consideration. From triangle inequality we have:

$$\left\| \frac{1}{\sqrt{n}} \sum_{i=0}^{n-1} \{h_{\theta_\star}(y_{0:i}) - h_{\theta_\star}(y_{-\infty:i})\} \right\|_{2+\delta} \leq \frac{1}{\sqrt{n}} \sum_{i=0}^{n-1} \|h_{\theta_\star}(y_{0:i}) - h_{\theta_\star}(y_{-\infty:i})\|_{2+\delta}.$$

Recall that $\nabla \ell_{\theta_\star}(y_{0:n-1}) = \sum_{i=0}^{n-1} h_{\theta_\star}(y_{0:i})$. Therefore recalling equation (5.22) and the definition

(5.24), the bound (5.27) implies:

$$\frac{\nabla_\theta \ell_{\theta_\star}(y_{0:n-1})}{\sqrt{n}} = \frac{1}{\sqrt{n}} \sum_{i=0}^{n-1} h_{\theta_\star}(y_{-\infty:i}) + \mathcal{O}_{L^{2+\delta}}\left(n^{-1/2}\right). \tag{5.28}$$

For $a \geq 1$ and a sequence of positive reals $\{b_k\}$, $\mathcal{O}_{L^a}$ denotes a sequence of random variables with L^a norm being $\mathcal{O}(b_n)$. Then we define the matrix as follows:

$$\mathcal{J}_{\theta_\star} := \mathbb{E}\left[h_{\theta_\star}(y_{-\infty:0}) h_{\theta_\star}(y_{-\infty:0})^\top\right]. \tag{5.29}$$

Step 4. At this point we need to make use of the model-correctness Assumption 8. Define the filtration $\mathcal{F}_i := \sigma(y_j; -\infty < j \leq i)$. Then one can show that $\{h_{\theta_\star}(y_{-\infty:i})\}_{i\geq 0}$ is strictly stationary, ergodic and square integrable martingale difference sequence as follows. Apparently $\{h_{\theta_\star}(y_{-\infty:i})\}_{i\geq 0}$ is adopted w.r.t. $\mathcal{F}_i$ and integrable by the assumptions. Then we consider the conditional expectation of $h_{\theta_\star}(y_{-\infty:i})$ given $\mathcal{F}_{i-1}$, that is:

$$\begin{aligned}
\mathbb{E}\left[h_{\theta_\star}(y_{-\infty:i}) \mid \mathcal{F}_{i-1}\right] = \mathbb{E}\left[\mathbb{E}\left[d_{\theta_\star}\left(x_{i-1,}x_i, y_i\right) \mid \mathcal{F}_i\right] \mid \mathcal{F}_{i-1}\right] \\
+ \sum_{j=-\infty}^{i-1} \mathbb{E}\left[\{\mathbb{E}\left[d_{\theta_\star}\left(x_{j-1,}x_j, y_j\right) \mid \mathcal{F}_i\right] \right. \\
\left. - \mathbb{E}\left[d_{\theta_*}\left(x_{j-1,}x_j, y_j\right) \mid \mathcal{F}_{i-1}\right]\right\} \mid \mathcal{F}_{i-1}\right].
\end{aligned}$$

Each term in the sum is trivially 0 by the Lebesgue's dominated convergence theorem . For the first term, we have:

$$\begin{aligned}
\mathbb{E}\left[d_{\theta_\star}\left(x_{i-1,}x_i, y_i\right) \mid \mathcal{F}_{i-1}\right] = \mathbb{E}\left[d_{\theta_*}\left(x_{i-1,}x_i, y_i\right) \mid x_{i-1}, \mathcal{F}_{i-1}\right] \\
\equiv 0.
\end{aligned}$$

Therefore, we conclude that $\mathbb{E}\left[h_{\theta_\star}(y_{-\infty:i}) \mid \mathcal{F}_{i-1}\right] = 0$. Note that since $\{y_k\}_{k\geq k}$ is strictly stationary and ergodic by Assumption 1, $\{h_{\theta_*}(y_{-\infty:i})\}_{i\geq 0}$ is also. Also we have indeed used the model correctness assumption to obtain the latter result. So, terms $h_{\theta_\star}(y_{-\infty:i})$ make up a strictly stationary, ergodic martingale increment sequence, of finite second moment, under the filtration generated by the data. We define:

$$M_{n,j} := \sum_{i=1}^{n-1} \left(h_{\theta_\star}(y_{-\infty:i})\right)_j, \ 1 \leq j \leq d.$$

Subscript j indicates the j-th component of the d-dimensional vectors. Then $\frac{1}{\sqrt{n}} M_n \Rightarrow \mathcal{N}_d(0, \mathcal{J}(\theta_*))$ follows from the martingale difference central limit theorem (see McLeish (1974) for instance) where the convergence is in distribution.

We now turn to the second term in (5.21).

Proposition 28. *Under Assumption 1 to Assumption 7, we have that, w.p.*1,

$$\lim_{\delta\to 0}\lim_{n\to\infty}\sup_{\theta\in\mathcal{B}_\delta(\theta_\star)}\left|(-\nabla\nabla^\top\ell_\theta(y_{0:n-1})/n)-\mathcal{J}_{\theta_\star}\right|=0.$$

Proof. This is Douc et al. (2014, Theorem 13.24). □

Proposition 29. *Under Assumption 1 to Assumption 7, we have that, w.p1.*

$$J_{\theta_\star}(y_{0:n-1}):=-\int_0^1\frac{1}{n}\nabla\nabla^\top\ell_{s\hat{\theta}_n+(1-s)\theta_\star}(y_{0:n-1})ds\to\mathcal{J}_{\theta_\star}.$$

Proof. This is implied immediately from Proposition 28 and (2) of Proposition 27. □

Notice that this result does not require the assumption of model correctness. We do make the following assumption on the HMM model under consideration.

Assumption 9. *The matrix $\mathcal{J}_{\theta_\star}\in\mathbb{R}^{d\times d}$ is non-singular.*

Thus, assuming n is big enough to permit inversion, a combination of equations (5.19) and (5.20)gives:

$$\ell_{\hat{\theta}_n}(y_{0:n-1})=\ell_{\theta_\star}(y_{0:n-1})+\frac{1}{2}\frac{\nabla^\top\ell_{\theta_\star}(y_{0:n-1})}{\sqrt{n}}J_{\theta_\star}(y_{0:n-1})^{-1}\frac{\nabla\ell_{\theta_\star}(y_{0:n-1})}{\sqrt{n}}.$$

We summarise the results in this part with the following proposition.

Proposition 30. *i) Under Assumption 1 to Assumption 7 and Assumption 9, we have that:*

$$\ell_{\hat{\theta}_n}(y_{0:n-1})=\ell_{\theta_\star}(y_{0:n-1})+\frac{1}{2}\frac{\nabla^\top\ell_{\theta_\star}(y_{0:n-1})}{\sqrt{n}}J_{\theta_*}(y_{0:n-1})^{-1}\frac{\nabla\ell_{\theta_\star}(y_{0:n-1})}{\sqrt{n}},$$

*whenever the inverse exists. Moreover $J_{\theta_\star}(y_{0:n-1})\to\mathcal{J}_{\theta_\star}$, w.p.*1*, for the non-singular matrix $\mathcal{J}_{\theta_\star}$ defined in (5.29).*

ii) Under Assumption 1 to Assumption 7 and Assumption 9, we have that:

$$\frac{1}{\sqrt{n}}\nabla\ell_{\theta_\star}(y_{0:n-1})=\frac{1}{\sqrt{n}}\sum_{i=1}^{n-1}h_{\theta_\star}(y_{-\infty:i})+\frac{1}{\sqrt{n}}R_n,$$

where $\|h_{\theta_*}(y_{-\infty:i})\|_{2+\delta}+\|R_n\|_{2+\delta}\le C$, for $\delta>0$ as determined in (3) of Assumption 7 and a constant $C>0$.

iii) Under Assumption 1 to Assumption 9. we obtain:

$$\nabla\ell_{\theta_\star}(y_{0:n-1})/\sqrt{n}\Rightarrow\mathcal{N}_d\left(0,\mathcal{J}_{\theta_\star}\right).$$

Proof. Parts (1) and (2) are simply rewritings of earlier calculations. (3) follows from combining the L^2-bound and the martingale difference central limit theorem, see Douc et al. (2014, Theorem 13.23). □

Theorem 21. *Suppose that Assumption 1 to Assumption 9 hold. Then we have that:*

i) $\limsup_{n\to\infty} \frac{|\nabla \ell_{\theta_\star}(y_{0:n-1})|}{\sqrt{2n\log\log n}} = \mathbb{E}\left[(h_{\theta_\star}(y_{-\infty:i}))_j^2\right]^{1/2},\ 1 \le j \le d.$

ii) $\ell_{\hat{\theta}_n}(y_{0:n-1}) = \ell_{\theta_\star}(y_{0:n-1}) + \mathcal{O}(\log\log n)$ *w.p.*1.

Proof. Proposition 30 implies that $\nabla \ell_{\theta_\star}(y_{0:n-1})$ can be decomposed into:

$$\nabla \ell_{\theta_\star}(y_{0:n-1}) = \sum_{i=1}^{n-1} h_{\theta_\star}(y_{-\infty:i}) + R_n,$$

where $\|h_{\theta_*}(y_{-\infty:i})\|_{2+\delta} + \|R_n\|_{2+\delta} \le C$, for $\delta > 0$ as determined in (3) of Assumption 7 and a constant $C > 0$. Recall that $h_{\theta_\star}(y_{-\infty:i})$ is strictly stationary, ergodic martingale increment sequence, of finite second moment, w.r.t. the filtration $\mathcal{F}_i := \sigma(y_j, -\infty < j \le i)$. To obtain the result, it is sufficient to show that for any $\epsilon > 0$ and $1 \le j \le d$, $\mathbb{P}\left(\limsup_{n\to\infty} \mid R_{n,j} \mid \ge \epsilon\sqrt{n}\right) = 0$. Using Markov inequality, we have that:

$$\begin{aligned}\mathbb{P}\left(|R_{n,j}| \ge \epsilon\sqrt{n}\right) = \mathbb{P}\left(|R_{n,j}|^{2+\delta} \ge \epsilon^{2+\delta} n^{1+\delta/2}\right) &\le \frac{\mathbb{E}\left[\mid R_{n,j} \mid^{2+\delta}\right]}{\epsilon^{2+\delta} n^{1+\delta/2}}, \\ &\le \frac{C}{\epsilon^{2+\delta} n^{1+\delta/2}}.\end{aligned}$$

Define the set $A_n := \{\mid R_{n,j} \mid \ge \epsilon\sqrt{n}\}$. Then above evaluation implies that $\sum_{n=0}^{\infty} \mathbb{P}(A_n) < \infty$ and this gives rise to $\mathbb{P}\left(\limsup_{n\to\infty} \mid R_{n,j} \mid \ge \epsilon\sqrt{n}\right) = 0$ by the Borel–Cantelli lemma. Then the LIL for the martingales (Stout, 1970) implies that:

$$\limsup_{n\to\infty} \frac{\mid \nabla_\theta \ell_{\theta_\star}(y_{0:n-1}) \mid}{\sqrt{2n\log\log n}} = \mathbb{E}\left[(h_{\theta_\star}(y_{-\infty:i}))_j^2\right]^{1/2},\ 1 \le j \le d.$$

The mean value theorem gives:

$$0 \equiv \frac{1}{n}\nabla \ell_{\hat{\theta}_n}(y_{0:n-1}) = \frac{1}{n}\nabla \ell_{\theta_\star}(y_{0:n-1}) - J_{\theta_\star}(y_{0:n-1})\left(\hat{\theta}_n - \theta_\star\right).$$

Above result implies that $\nabla \ell_{\theta_\star}(y_{0:n-1}) = \mathcal{O}\left(\sqrt{n\log\log n}\right)$ w.p.1. In addition, Proposition 29 gives rise to $J_{\theta_*}(y_{0:n-1}) = \mathcal{O}(1)$ w.p.1. Therefore, Solving for $\left(\hat{\theta}_n - \theta_*\right)$ gives:

$$\begin{aligned}\hat{\theta}_n - \theta_\star &= J_{\theta_\star}(y_{0:n-1})^{-1}\left(\frac{1}{n}\nabla \ell_{\theta_\star}(y_{0:n-1})\right), \\ &= \mathcal{O}\left(\sqrt{n^{-1}\log\log n}\right),\end{aligned}$$

holds w.p.1. In the same manner, applying the mean value theorem to $\ell_{\hat{\theta}_n}(y_{0:n-1})$ gives rise to, w.p.1,

$$\begin{aligned}\ell_{\hat{\theta}_n}(y_{0:n-1}) - \ell_{\theta_\star}(y_{0:n-1}) &= \left(\hat{\theta}_n - \theta_\star\right)^\top \nabla \ell_{\theta_\star}(y_{0:n-1}) \\ &\quad + \frac{1}{2}\sqrt{n}\left(\hat{\theta}_n - \theta_\star\right)^\top J_{\theta_\star}(y_{0:n-1})\sqrt{n}\left(\hat{\theta}_n - \theta_\star\right), \\ &= \mathcal{O}(\log\log n).\end{aligned}$$

□

5.5 Model selection criteria for HMMs

We provide a brief illustration for the derivation of AIC and BIC, with focus on HMMs. Results obtained that explicitly connect BIC and the evidence will allow for deriving consistency properties for the evidence directly after studying the BIC criterion later in the paper.

5.5.1 BIC and evidence for HMMs

We consider the derivation of BIC for general HMMs. BIC is proposed by Schwarz (1978) and can be obtained by applying the Laplace approximation to the marginal likelihood (or evidence) of the model under consideration. Consideration of the sequence of log-likelihood functions over the sample size n (see e.g. Kass et al. (1990) for the concept of 'Laplace-regular' models) provide sufficient analytical conditions for controlling the difference between the evidence and BIC. We briefly review the Taylor expansions underlying the derivation of BIC and provide the regularity conditions that control its difference from the evidence in the context of HMMs. Compared with Kass et al. (1990), weaker conditions are required here, as BIC derives from an $\mathcal{O}\left(n^{-1}\right)$ approximation of the evidence (rather than $\mathcal{O}\left(n^{-2}\right)$ expansions looked at in the Laplace-regular framework).

Let $\pi(\theta)$ be the prior density w.r.t. an appropriate reference measure (e.g. the Lebesgue measure on $\mathbb{R}^d$) $d\theta$ for the parameter θ. Then the evidence $m(y_{0:n-1})$ is given by:

$$m(y_{0:n-1}) = \int_{\Theta} \pi(\theta) \exp\left\{\ell_\theta(y_{0:n-1})\right\} d\theta. \tag{5.30}$$

We define:

$$\mathsf{J}_{\hat{\theta}_n}(y_{0:n-1}) := -\frac{1}{n}\nabla\nabla^\top \ell_{\hat{\theta}_n}(y_{0:n-1}).$$

We will be explicit on regularity conditions in the statement of the Proposition that follows. Following similar steps as in Schervish (2012); Kass et al. (1990), we apply a fourth-order Taylor expansion around the MLE θ_n that gives, for $u := \sqrt{n}(\theta - \hat{\theta}_n)$:

$$\begin{aligned}\ell_\theta(y_{0:n-1}) = \ell_{\hat{\theta}_n}(y_{0:n-1}) - \frac{1}{2}u^\top \mathsf{J}_{\hat{\theta}_n}(y_{0:n-1})u \\ +\frac{1}{6}n^{-1/2}\sum_{i,j,k=1}^{d} u_i u_j u_k \frac{\partial_{\theta_i}\partial_{\theta_j}\partial_{\theta_k}\ell_{\hat{\theta}_n}(y_{0:n-1})}{n} + R_{1,n},\end{aligned} \tag{5.31}$$

for the residual term $R_{1,n}$ involving fourth-order derivatives of the log-likelihood $\theta \mapsto \ell_\theta(y_{0:n-1})$ evaluated at $\xi := a\hat{\theta}_n + (1-a)\theta$ for some $a \in [0,1]$, fourth order polynomials of u, and a factor of n^{-1}, see e.g. Lang (2012, Chapter 14) and Appendix A for details on such expansions. Notice that we have used $\nabla\ell_{\hat{\theta}_n}(y_{0:n-1}) = 0$. For the prior density, we have:

$$\pi(\theta) = \pi(\hat{\theta}_n) + n^{-1/2}\nabla^\top\pi(\hat{\theta}_n)u + R_{2,n},$$

for the residual $R_{2,n}$ with second-order derivatives of $\pi(\theta)$, second-order polynomial of u and a factor of n^{-1}. Using a second order expansion for $x \mapsto e^x$, only for the the terms beyond the quadratic in u in (5.31), we obtain:

$$\frac{m(y_{0:n-1})}{p_{\hat{\theta}_n}(y_{0:n-1})} = \int_\Theta e^{-\frac{1}{2}u^\top \mathsf{J}_{\hat{\theta}}(y_{0:n-1})u} \times \left\{\pi(\hat{\theta}_n) + n^{-1/2}\mathcal{H}(u,\, y_{0:n-1}) + R_n\right\} d\theta, \tag{5.32}$$

where we have separated the term (later on removed as having zero mean under a Gaussian integrator Schervish (2012)):

$$\mathcal{H}(u, y_{0:n-1}) := \frac{1}{6}n^{-1/2}\sum_{i,j,k=1}^{d} u_i u_j u_k \frac{\partial_{\theta_i}\partial_{\theta_j}\partial_{\theta_k}\ell_{\hat{\theta}_n}(y_{0:n-1})}{n} + \nabla^\top \pi(\hat{\theta}_n)u;$$

the residual term R_n can be deduced from the calculations.

Remark 9. *The Laplace-regular setting of Kass et al. (1990) provides concrete conditions for the above derivations to be valid and for controlling the deduced residual terms. Apart from the standard assumptions on the existence of derivatives and a bound on fourth order derivative of close to θ_* - the latter being defined in Proposition 27 as the limit of $\hat{\theta}_n$ - the following are also required*:

i) *For any $\delta > 0$, w.p.1:*

$$\limsup_{n\to\infty} \sup_{\theta\in\Theta\setminus\mathcal{B}_\delta(\theta_*)} \left\{\frac{1}{n}\left(\ell_\theta(y_{0:n-1}) - \ell_{\theta_*}(y_{0:n-1})\right)\right\} < 0;$$

ii) *For any $\epsilon > 0$, $\mathcal{B}_\epsilon(\theta_*) \subseteq \Theta$, and w.p.1:*

$$\limsup_{n\to\infty} \sup_{\theta\in\mathcal{B}_\epsilon(\theta_*)} \left\{\frac{1}{n}\det\left(\nabla\nabla^\top \ell_\theta(y_{0:n-1})\right)\right\} < 0.$$

Note that (1) *is implied by Proposition 27 and identifiability Assumption 6. Also, Proposition 28 implies* (2).

Here, $\det(\cdot)$ denotes the determinant of a square matrix. Following the above remark, the Laplace-regular setting of Kass et al. (1990) translate into the following assumption and proposition.

Assumption 10. *i) (The log-likelihood function $\theta \mapsto \ell_\theta(y_{0:n-1})$ is four-times continuously differentiable in θ w.p.*1. *Also the prior $\theta \mapsto \pi(\theta)$ is two-times continuously differentiable w.p.1.*

ii) For some $\epsilon > 0$, $\mathcal{B}_\epsilon(\theta_*) \subseteq \Theta$ and w.p.1, for all $1 \le j_1 \cdots \le j_k \le d$, with $0 \le k \le 4$, ,

$$\limsup_{n\to\infty} \sup_{\theta\in\mathcal{B}_\epsilon(\theta_*)} \left\{\frac{1}{n}\left|\frac{\partial^{j_1\cdots j_k}\ell_\theta(y_{0:n-1})}{\partial\theta_{j_1}\cdots\partial\theta_{j_k}}\right|\right\} < \infty,$$

holds w.p.1.

Proposition 31. *Under Assumption 1 to Assumption 7, Assumption 9 and Assumption 10, we have that, w.p.1:*

$$\frac{m(y_{0:n-1})}{p_{\hat{\theta}_n}(y_{0:n-1})} = (2\pi)^{d/2} n^{-d/2} \left[\det\left(\mathsf{J}_{\hat{\theta}_n}(y_{0:n-1})\right)\right]^{-1/2} \pi(\hat{\theta}_n)\left(1+\mathcal{O}\left(n^{-1}\right)\right).$$

Proof. Under the assumptions, Tadic and Doucet (2018, Theorem 2.1) ensure that the the log-likelihood $\theta \mapsto \ell_\theta(y_{0:n-1})$ is four-times continuously differentiable. Recall from Proposition 27 that $\lim_{n\to\infty} \sup_{\theta\in\Theta} \left|\frac{1}{n}\ell_\theta(y_{0:n-1}) - \ell(\theta)\right| = 0$ *w.p.*1 and $\hat{\theta}_n \to \theta_*$ *w.p.*1. Note that θ_* is the unique maximiser of $\ell(\cdot)$, under Assumption 6. Then we choose sufficiently small $\delta > 0$ (in (1) of Remark 1) and $\gamma > 0$ such that for large enough n, $\mathcal{B}_\delta(\theta_*) \subseteq \mathcal{B}_\gamma(\hat{\theta}_n) \subseteq \mathcal{B}_{\min\{\epsilon_1\epsilon_2\}}(\theta_*)$ with $\epsilon = \epsilon_1$ and $\epsilon = \epsilon_2$ in (2) of Assumption 10 and (2) of Remark 9 respectively. As a result, we have that:

$$\begin{aligned}
\frac{m(y_{0:n-1})}{p_{\hat{\theta}_n}(y_{0:n-1})} &= \int_{\Theta\setminus\mathcal{B}_\gamma(\hat{\theta}_n)} \pi(\theta)\exp\left(n\times\frac{1}{n}\left\{\ell_\theta(y_{0:n-1}) - \ell_{\hat{\theta}_n}(y_{0:n-1})\right\}\right)d\theta \\
&\quad + \int_{\mathcal{B}_\gamma(\hat{\theta}_n)} \pi(\theta)\exp\left(\left\{\ell_\theta(y_{0:n-1}) - \ell_{\hat{\theta}_n}(y_{0:n-1})\right\}\right)d\theta \\
&\leq \exp(-cn) + \int_{\mathcal{B}_\gamma(\hat{\theta}_n)} \pi(\theta)\exp\left(\left\{\ell_\theta(y_{0:n-1}) - \ell_{\hat{\theta}_n}(y_{0:n-1})\right\}\right)d\theta,
\end{aligned}$$

for some constant $c > 0$, where we used (1) of Remark 9 to obtain the inequality. It remains to treat the integral on $\mathcal{B}_\gamma(\hat{\theta}_n)$. Application of the Taylor expansions as we have described and continuing from (5.32), with the domain of integration now being $\mathcal{B}_\gamma(\hat{\theta}_n)$, will yield, *w.p.*1,

$$\begin{aligned}
\mathcal{I}_n &:= \int_{\mathcal{B}_\gamma(\hat{\theta}_n)} \pi(\theta)\exp\left(\left\{\ell_\theta(y_{0:n-1}) - \ell_{\hat{\theta}_n}(y_{0:n-1})\right\}\right)d\theta \\
&= \int_{\mathcal{B}_\gamma(\hat{\theta}_n)} \exp\left(-\frac{1}{2}u^\top \mathsf{J}_{\hat{\theta}_n}(y_{0:n-1})\right)\left\{\pi(\hat{\theta}_n) + n^{-1/2}\mathcal{H}(u, y_{0:n-1}) + R_n\right\}d\theta. \qquad (5.33)
\end{aligned}$$

A careful, but otherwise straightforward, consideration of the structure of the residual R_n, gives that, w.p.1:

$$\begin{aligned}
&\frac{1}{(2\pi)^{d/2}\left\{\det\left(\mathsf{J}_{\hat{\theta}_n}(y_{0:n-1})\right)\right\}^{-1/2}} \int_{\mathcal{B}_\gamma(\hat{\theta}_n)} \exp\left(-\frac{1}{2}u^\top \mathsf{J}_{\hat{\theta}_n}(y_{0:n-1})\right) \mid R_n \mid d\theta \\
&\qquad\qquad = \mathcal{O}(n^{-1}),
\end{aligned}$$

where we used Remark 1 and (2) of Assumption 9. Therefore, continuing from (5.33), the change of variables $u = \sqrt{n}(\theta - \hat{\theta}_n)$ implies that, for $f(\cdot;\Omega)$ denoting the pdf of a centred Gaussian distribution

with precision matrix Ω,

$$\begin{aligned}\mathcal{I}_n = (2\pi)^{d/2} \left\{\det\left(\mathsf{J}_{\hat{\theta}_n}(y_{0:n-1})\right)\right\}^{-1/2} \\ \times \int_{\mathcal{B}_{\gamma\sqrt{n}}(0)} f(u; \mathsf{J}_{\hat{\theta}_n}(y_{0:n-1})) \left\{\pi(\hat{\theta}_n) + n^{-1/2}\mathcal{H}(u, y_{0:n-1})\right\} du \\ \times (1 + \mathcal{O}(n^{-1})),\end{aligned}$$

w.p.1. The final result follows from the fact, that using (1) of Assumption 10, the integral appearing above is $\mathcal{O}(\exp(-c'n))$ apart from the same integral over the whole $\mathbb{R}^d$, for some constant $c' > 0$. □

Therefore, Proposition 31 implies that:

$$\begin{aligned}m(y_{0:n-1}) = \exp\left(n\frac{1}{n}\ell_{\hat{\theta}_n}(y_{0:n-1})\right) \frac{(2\pi)^{d/2}\,\pi(\hat{\theta}_n)}{n^{d/2}\left[\det\left(\mathsf{J}_{\hat{\theta}_n}(y_{0:n-1})\right)\right]} \\ \times \left(1 + \mathcal{O}(n^{-1})\right),\end{aligned}$$

w.p1. Then taking logarithm of both sides gives rise to:

$$\begin{aligned}-2\log m(y_{0:n-1}) = -2\ell_{\hat{\theta}_n}(y_{0:n-1}) - 2\log\pi\left(\hat{\theta}_n\right) - d\log 2\pi + d\log n \\ + \log\det\left(\mathsf{J}_{\hat{\theta}_n}(y_{0:n-1})\right) + \mathcal{O}(n^{-1}),\ w.p.1.\end{aligned}$$

Note that $\pi(\hat{\theta}_n) \to \pi(\theta_*)$ as $n \to \infty$ holds w.p.1 by the assumption, so this term will be $\mathcal{O}(1)$. Also (2) of Proposition 30 implies that $\mathsf{J}_{\hat{\theta}_n}(y_{0:n-1}) = \mathcal{O}(1)$, w.p.1, thus, the continuous mapping theorem gives rises to $\log\det\left(\mathsf{J}_{\hat{\theta}_n}(y_{0:n-1})\right) = \mathcal{O}(1)$, *w.p.*1. As a consequence, ignoring $\mathcal{O}(1)$ terms w.r.t. n yields:

$$-2\log m(y_{0:n-1}) \approx -2\ell_{\hat{\theta}_n}(y_{0:n-1}) + d\log n,\ w.p.1.$$

Thus, working with the Laplace approximation to the marginal likelihood , one can derive BIC:

$$BIC := -2\ell_{\hat{\theta}_n}(y_{0:n-1}) + d\log n. \tag{5.34}$$

Remark 10. *The above results provide a significant conceptual reassurance. Admitting the evidence as the core principle under which model comparison is carried out, if amongst a family of parametric HMM models one has the largest evidence for any big enough n w.p.1, then BIC is guaranteed to select that model as the optimal one eventually since they are asymptotically equivalent.*

Remark 11. *There is considerable work in the literature regarding consistency properties of the evidence (or Bayes Factor) for classes of models beyond the i.i.d. setting, see e.g. Chatterjee et al. (2020) and the references therein. In our approach, we have brought together results in the literature to deliver assumptions that – whilst being fairly general – were produced with HMMs in mind (and the connection between AIC and the evidence) and are relatively straightforward to be verified, indeed, for HMMs. Alternative approaches typically provide higher-level conditions (see e.g. above reference) in an attempt to preserve generality*

5.5.2 AIC for HMMs

AIC is developed in Akaike (1974) where its derivation is discussed for i.i.d. data and Gaussian models of ARMA type. Following more recent expositions (see e.g. Claeskens and Hjort (2008)), AIC is based on the use of the Kullback-Leibler (KL) divergence for quantifying the distance between the true data-generating distribution and the probability model; an effort to reduce the bias of a 'naive' estimator of the KL divergence leads to the formula for AIC. The case that one does not assume that the parametric model contains the true data distribution corresponds to a generalised version of AIC often called the Takeuchi Information Criterion (TIC), first proposed in Takeuchi (1976). The above ideas are easy to be demonstrated in simple settings (e.g. Claeskens and Hjort (2008) consider i.i.d. and linear regression models).

The framework connecting KL with AIC, in the context of HMMs, can be developed as follows. Let $p_\star(dy_{0:n-1})$ denote the true data-generating distribution, $n \geq 1$. A model is suggested in the form of a family of distributions $\{p_\theta(dy_{0:n-1});\, \theta \in \Theta\}$. We assume that $p_\star(dy_{0:n-1})$ and $p_\theta(dy_{0:n-1})$ admit the densities $p_\star(y_{0:n-1})$, $p_\theta(y_{0:n-1})$ *w.r.t.* $\nu^{\otimes n}$, $n \geq 1$. We work with the KL distance:

$$\begin{aligned} KL_n(\theta) &:= \frac{1}{n} \int p_\star(dz_{0:n-1}) \log \frac{p_\star(z_{0:n-1})}{p_\theta(z_{0:n-1})}, \\ &= \frac{1}{n} \int p_\star(dz_{0:n-1}) \log p_\star(z_{0:n-1}) - \frac{1}{n} \int p_\star(dz_{0:n-1}) \log p_\theta(y_{0:n-1}). \end{aligned} \tag{5.35}$$

Therefore, minimizing $KL_n(\theta)$ is equivalent to maximizing:

$$\mathcal{R}_n(\theta) := \frac{1}{n} \int p_\star(dy_{0:n-1}) \log p_\theta(y_{0:n-1}).$$

Following standard ideas from cases models (e.g. i.i.d. models), one is interested in the quantity $\mathcal{R}_n(\hat{\theta}_n)$, but, in practice, only has access to the naive estimator $\frac{1}{n}\ell_{\hat{\theta}_n}(y_{0:n-1})$, the latter tending to have positive bias versus $\mathcal{R}_n(\hat{\theta}_n)$ due to the use of both the data and the data-induced MLE in its expression. AIC is then derived by finding the larger order term (of size $\mathcal{O}(1/n)$) in the discrepancy of the expectation and appropriately adjusting the naive estimator. We make the following assumptions.

Assumption 11. *i) There exists constant $C > 0$, such that w.p.*1,

$$\sup_{n \geq 1} \sup_{\theta \in \Theta} \left\{ \frac{1}{n} \mid \nabla \nabla^\top \ell_\theta(y_{0:n-1}) \mid \right\} < C.$$

*ii) There is some $n_0 > 1$ such that w.p.*1*, matrix $J_{\theta_*}^{-1}(y_{0:n-1})$ defined in Proposition 29, is well-posed for all $n > n_0$, and there is a constant $C' > 0$, such that w.p.*1,

$$\sup_{n > n_0} \mid J_{\theta_*}^{-1}(y_{0:n-1}) \mid < C'.$$

These are high-level assumptions, especially (2) of Assumption 11, and a more analytical study is required for them to be of immediate practical use (or weakening them); but such a study would considerably deviate from the main purposes of this work. Our contribution is contained in the

following Proposition.

Proposition 32. *Under Assumption 1 to Assumption 9 and Assumption 11, we have that:*

$$\mathbb{E}\left[\frac{1}{n}\ell_{\hat{\theta}_n}(y_{0:n-1}) - \mathcal{R}_n(\hat{\theta}_n)\right] = \frac{d}{n} + o(n^{-1}).$$

Proof. Applying a second order Taylor expansion around $\theta_\star$ yields:

$$\begin{aligned}
&\frac{1}{n}\ell_{\hat{\theta}_n}(y_{0:n-1}) - \mathcal{R}_n(\hat{\theta}_n) \\
=&\frac{1}{n}\ell_{\theta_\star}(y_{0:n-1}) - \frac{1}{n}\int \ell_{\theta_\star}(z_{0:n-1})p_\star(dz_{0:n-1}) \\
&+ \frac{1}{n}\frac{\nabla^\top \ell_{\theta_\star}(y_{0:n-1})}{\sqrt{n}}\sqrt{n}(\hat{\theta}_n - \theta_\star) - \left\{\int \nabla^\top \ell_{\theta_\star}(z_{0:n-1})p_\star(dz_{0:n-1})\right\}(\hat{\theta}_n - \theta_\star) \\
&+ \frac{1}{2n}\sqrt{n}(\hat{\theta}_n - \theta_\star)^\top \left\{\int \mathcal{E}_{\theta_\star}(y_{0:n-1}, z_{0:n-1})p_\star(dz_{0:n-1})\right\}\sqrt{n}(\hat{\theta}_n - \theta_\star),
\end{aligned} \tag{5.36}$$

where we have set:

$$\mathcal{E}_{\theta_\star}(y_{0:n-1}, z_{0:n-1}) := \int_0^1 \frac{\nabla\nabla^\top \ell_{s\hat{\theta}_n + (1-s)\theta_\star}(y_{0:n-1})ds - \nabla\nabla^\top \ell_{s\hat{\theta}_n + (1-s)\theta_\star}(z_{0:n-1})}{n} ds.$$

Taking expectations in (5.36), notice that: 1) the expectation of the first difference on the right-hand-side is trivially 0; 2) the integral appearing in the second difference in identically zero, since we are working under the correct model Assumption 8. It remains to consider the expectation of the terms:

$$\begin{aligned}
\zeta_n &:= \frac{1}{n}\frac{\nabla^\top \ell_{\theta_\star}(y_{0:n-1})}{\sqrt{n}}\sqrt{n}(\hat{\theta}_n - \theta_\star), \\
\zeta_n^{'} &:= \frac{1}{2n}\sqrt{n}(\hat{\theta}_n - \theta_\star)^\top \left\{\int \mathcal{E}_{\theta_\star}(y_{0:n-1}, z_{0:n-1})p_\star(dz_{0:n-1})\right\}\sqrt{n}(\hat{\theta}_n - \theta_\star).
\end{aligned} \tag{5.37}$$

First, ζ_n can be expressed as:

$$\zeta_n = \frac{1}{n} \times \frac{\nabla^\top \ell_{\theta_\star}(y_{0:n-1})}{\sqrt{n}} J_{\theta_\star}(y_{0:n-1})^{-1} \frac{\nabla \ell_{\theta_\star}(y_{0:n-1})}{\sqrt{n}}.$$

Thus, Proposition 30 and Slutsky's theorem give that:

$$n\zeta_n \Rightarrow Z^\top \mathcal{J}_{\theta_\star}^{-1} Z;\ Z \sim \mathcal{N}(0, \mathcal{J}_{\theta_\star}),$$

where $\Rightarrow$ denotes weak convergence. For weak convergence to imply convergence in expectation, we require uniform integrability. (2) of Assumption 11 takes care of the difficult term $J_{\theta_\star}(y_{0:n-1})^{-1}$, see Appendix D . Then, Proposition 30 and the Marcinkiewicz-Zygmund inequality can be applied for martingales Ibragimov and Sharakhmetov (1999), give that:

$$\sup_n \left\| \frac{\nabla \ell_{\theta_\star}(y_{0:n-1})}{\sqrt{n}} \right\|_2 < \infty.$$

Thus, from Cauchy-Schwarz, we have:

$$\sup \|n\zeta_n\|_2 < \infty,$$

which implies uniform integrability for $\{n\zeta_n\}_n$. So, we have shown that:

$$\mathbb{E}\left[n\zeta_n\right] \to \mathbb{E}\left[Z^\top \mathcal{J}_{\theta_\star}^{-1} Z\right] \equiv d. \tag{5.38}$$

We proceed to term $\zeta_n^{'}$ in (5.37). Define $A_{\theta_\star}(y_{0:n-1}) := \nabla^\top \ell_{\theta_\star}(y_{0:n-1})/\sqrt{n} \times J_{\theta_\star}^{-1}(y_{0:n-1})$, and we have that:

$$2n\zeta_n^{'} = A_{\theta_\star}(y_{0:n-1}) \left\{ \int \mathcal{E}_{\theta_\star}(y_{0:n-1}, z_{0:n-1}) p_\star(dz_{0:n-1}) \right\} A_{\theta_\star}^\top(y_{0:n-1}).$$

Clearly, we can write:

$$\mathbb{E}\left[2n\zeta_n^{'}\right] = \int \left\{ A_{\theta_\star}(y_{0:n-1}) \mathcal{E}_{\theta_\star}(y_{0:n-1}, z_{0:n-1}) p_\star(dz_{0:n-1}) A_{\theta_\star}^\top(y_{0:n-1}) \right\} (\nu \otimes \nu)(dy_{0:n-1}, dz_{0:n-1}).$$

From Proposition 28 we obtain that $(\nu \otimes \nu)(dy_{0:n-1}, dz_{0:n-1})$-$w.p.1$, we have that $\lim_{n\to\infty} \mathcal{E}_{\theta_\star}(y_{0:n-1}, z_{0:n-1}) = \mathcal{J}_{\theta_\star} - \mathcal{J}_{\theta_\star} = 0$. This implies the weak convergence of $A_{\theta_\star}(y_{0:n-1}) \mathcal{E}_{\theta_\star}(y_{0:n-1}, z_{0:n-1}) p_\star(dz_{0:n-1}) A_{\theta_\star}^\top(y_{0:n-1}) \Rightarrow 0$. Assumption 11, and arguments similar to the ones used for $n\zeta_n$, imply uniform integrability for $n\zeta_n^{'}$. We thus have $\mathbb{E}\left[n\zeta_n^{'}\right] \to 0$. This latter result together with (5.38) completes the proof. □

Proposition 32 provides the underlying principle for use of the standard AIC:

$$AIC := -2\ell_{\hat{\theta}_n}(y_{0:n-1}) + 2d. \tag{5.39}$$

Remark 12. *We note that the only difference between BIC and AIC is the penalty term, that is, $d \log n$ for BIC and $2d$ for AIC. The penalty term for BIC grows to infinity as $n \to \infty$ in contrast with one of AIC which does not depend on the sample size n. More precisely, for all $n \geq 8$, BIC more heavily penalises a model with many parameters. As we will confirm in Section 5, this difference makes the significant difference regarding the consistency of model selection.*

5.6 BIC, evidence, AIC consistency properties

We will now use the results we have developed to examine the asymptotic properties of BIC, the evidence and AIC in the context of HMMs. We define the notions of strong and weak consistency in model selection in a nested setting as follows.

Definition 31. (Consistency of Model Selection Criterion). *Assume a sequence of nested parametric models*

$$\mathcal{M}_1 \subset \cdots \mathcal{M}_k \subset \cdots \mathcal{M}_p,\ k \geq 1,$$

specified via a sequence of corresponding parameter spaces $\Theta^1 \subseteq \mathbb{R}^{d_1}$, and $\Theta^{k+1} = \Theta^k \times \Delta\Theta^k$, $\Delta\Theta^k \subseteq \mathbb{R}^{d_k+1}$, $k \geq 1$ with $d_k < d_l$ for k < l. Let $\mathcal{M}_{k^\star}$, for some $k^\star \geq 1$, be the smallest model containing the

'correct' one the latter corresponding to the parameter value $\theta^{k^\star}(=\theta_\star) \in \Theta^{k^\star}$.

Let $\mathcal{M}_{\hat{k}_n}$, for index $\hat{k}_n$ based on data $y_{0:n-1} \in \mathsf{Y}^n$, $n \geqq 1$ be the model selected via optimising a Model Selection Criterion. If it holds that $\hat{k}_n \to k^\star$ as $n \to \infty$, w.p1, then the the Model Selection Criterion is called *strongly consistent*. If it holds that $\hat{k}_n \to k^\star$ as $n \to \infty$, in probability, then the Model Selection Criterion is said to be *weakly consistent*.

Assumption 12. *Assumption 1 to Assumption 6 hold for all parametric models $\mathcal{M}_k$, for index $1 \leq k < k^\star$; Assumption 1 to Assumption 9 hold for all parametric models $\mathcal{M}_k$, for index $k^\star \leq k \leq p$.*

We henceforth assume that for each $1 \leq k \leq p$, $\mathcal{M}_k$ corresponds to a parametric HMM as defined in Section 2. Also we will use the notation d_k to denote the dimension of a model $\mathcal{M}_k$. The particular model under consideration will be implied by the corresponding parameter appearing in an expression; i.e., a quantity involving θ^k will refer to model $\mathcal{M}_k$. E.g., $\theta^k_\star \in \Theta^k$ is the a.s. limit of the MLE, $\hat{\theta}^k_n$, for the model $\mathcal{M}_k$, and such a limit has been shown to exist under Assumption 1 to Assumption 6 for model model $\mathcal{M}_k$.

Remark 13. *For a model $\mathcal{M}_k$ that contains $\mathcal{M}_{k^\star}$ $(k > k^\star)$ for all of Assumption 2 to Assumption 9 to hold, it is necessary that the parametrisation of the larger model $\mathcal{M}_k$ is such that non-identifiability issues are avoided. In a trivial example, for $\mathcal{M}_{k^\star}$ corresponding to i.i.d. data from $\mathcal{N}(\theta_1, 1)$, a larger model of the form would satisfy Assumption 2 and Assumption 9 (the main ones the relate to the shape, in the limit, of the log-likelihood and, consequently identifiability) –one can check this – whereas model $\mathcal{N}(\theta_1 + \theta_2, 1)$ would not. In practice, for a given application with nested models, one can most times easily deduce whether identifiability issues are taken care of, thus Assumption 2 to Assumption 9 correspond to reasonable requirements over the larger models. In general, only 'atypical' parameterisations can produce non-identifyibility issues, thus, also abnormal behavior of the log-likelihood function , for the case of the larger model.*

Proposition 33. *Define $\lambda_n := \ell_{\hat{\theta}^k_n}(y_{0:n-1}) - \ell_{\hat{\theta}^{k^\star}_n}(y_{0:n-1})$, for $k \neq k^\star$. Assume that Assumption 1 to Assumption 9 and Assumption 12 hold. Then we have:*

i) If $\mathcal{M}_k \subset \mathcal{M}_{k^\star}$, then $\lim_{n\to\infty} n^{-1}\lambda_n \to \ell(\theta^k_\star) - \ell(\theta^{k^\star}_\star) < 0$, w.p1.

ii) If $\mathcal{M}_k \supset \mathcal{M}_{k^\star}$ then $\lambda_n > 0$ and $\lambda_n = \mathcal{O}(\log\log n)$, w.p.1.

Proof. From (1) of the Proposition 27 we have, w.p.1:

$$\begin{aligned} n^{-1}\left(\ell_{\hat{\theta}^k_n}(y_{0:n-1}) - \ell_{\hat{\theta}^{k^\star}_n}(y_{0:n-1})\right) &\to \ell(\theta^k_\star) - \ell(\theta^{k^\star}_\star), \\ \equiv \mathbb{E}\left[\log p_{\theta^k_\star}(y_0 \mid y_{-\infty:-1})\right] - \mathbb{E}\left[\log p_{\theta^{k^\star}_\star}(y_0 \mid y_{-\infty:-1})\right]&. \end{aligned}$$

Using Jensen's inequality and simple calculations, one obtains that:

$$\begin{aligned} &\mathbb{E}\left[\log p_{\theta^k_\star}(y_0 \mid y_{-\infty:-1})\right] - \mathbb{E}\left[\log p_{\theta^{k^\star}_\star}(y_0 \mid y_{-\infty:-1})\right] \\ &\leq \log \int \frac{p_{\theta^k_\star}(y_0 \mid y_{-\infty:-1})}{p_{\theta^{k^\star}_\star}(y_0 \mid y_{-\infty:-1})} p_{\theta^{k^\star}_\star}(y_{-\infty:0}) dy_{-\infty:0} \equiv \log 1. \end{aligned}$$

For strict inequality, Assumption 6 and Assumption 12 imply that mapping $\theta \mapsto \ell(\theta_\star^{k^\star})$ has the unique maximum $\theta_\star^{k^\star} \in \Theta_{k^\star}$. Thus, we cannot have $\ell(\theta_\star^k) = \ell(\theta_\star^{k^\star})$, as this would give (from the nested model structure) $\ell(\theta_\star^{k^\star}) = \ell(\theta_k, \theta_0)$ for some $\theta_0 \in \prod_{l=k+1}^{k^\star} \Delta\Theta_l$,with $(\theta_k, \theta_0)^\top \neq \theta_\star^{k^\star}$ (otherwise the definition of correct model class would be violated).

Having $\lambda_n \geq 0$ is a consequence of the log-likelihood for model $\mathcal{M}_k$ being maximised over a larger parameter domain than $\mathcal{M}_{k^\star}$. Then, notice that the limiting matrix $\mathcal{J}_{\theta_\star}$ in Proposition 29 (for the notation used therein) is positive-definite: it is non-negative-definite following its definition; then, non-singularity Assumption 9 provides the positive-definiteness. From Proposition 30, the difference in the definition of λ_n equals the difference of two quadratic forms, as the constants in the expression for the log-likelihood provided by Proposition 30 are equal for models $\mathcal{M}_{k^\star}$ and $\mathcal{M}_k$ cancel out. As $\lambda_n \geq 0$, and both quadratic forms are non-negative, it suffices to consider the one for model $\mathcal{M}_k$. The a.s. convergence of the positive-definite matrix in the quadratic form implies a.s. convergence of its eigenvalues and eigenvectors. Thus, using Theorem 21, overall one has that $\lambda_n = \mathcal{O}(\sum_{i=1}^{d} \log\log n) = \mathcal{O}(\log\log n)$. □

5.6.1 Asymptotic properties of BIC and evidence

BIC in known to be strongly consistent in i.i.d. settings and some particular non-i.i.d. ones, e.g. Claeskens and Hjort (2008). In the case of HMMs, Gassiat and Boucheron (2003) shows the strong consistency of BIC when the observations take a finite set of values. We prove the strong consistency of BIC and evidence for general HMMs. The key tool to obtain strong consistency of BIC in a general HMM is LIL we have obtained. Nishii (1988) also uses LIL for the i.i.d. setting to prove strong (and weak) consistency of BIC.

Recall that $k^\star$ denotes the index of the correct model.

Proposition 34. *i) Let $\hat{k}_n$ be the index of the model selected via minimizing BIC in (5.34) and k^* be the true one. Then, under Assumption 1 to Assumption 8 and Assumption 10, we have that $\hat{k}_n \to k^*$, w.p.1.*

ii) If $\hat{\mathsf{k}}_n$ denotes the index obtained via maximising the evidence in (5.30), then Assumption 1 to Assumption 10 imply that $\hat{\mathsf{k}}_n \to k^*$, w.p.1.

Proof. (1) We make use of Proposition 33. Notice that, since the model $\hat{k}_n$ attains the minimum BIC, $BIC_n\left(\mathcal{M}_{\hat{k}}\right) \leq BIC_n\left(\mathcal{M}_{k^*}\right)$ holds for large enough n. Fix the model $\mathcal{M}_k$. In the case that$\mathcal{M}_k \supset \mathcal{M}_{k^*}$, by (2) of Proposition 33 we have:

$$\begin{aligned}
BIC_n\left(\mathcal{M}_k\right) - BIC_n\left(\mathcal{M}_{k^*}\right) &= (d_k - d_{k^*}) \log n \\
&\quad - \left\{ \ell_{\hat{\theta}_n^k}(y_{0:n-1}) - \ell_{\hat{\theta}_n^{k^*}}(y_{0:n-1}) \right\}, \\
&= (d_k - d_{k^*}) \log n - \mathcal{O}(\log\log n), \\
&= \log\log n \left\{ \frac{(d_k - d_{k^*}) \log n}{\log\log n} - \mathcal{O}(1) \right\}, \\
&\to +\infty \; w.p.1,
\end{aligned}$$

since by the assumption $d_k - d_{k^*} > 0$ and thus $\frac{(d_k - d_{k^*})\log n}{\log\log n} \to \infty$. Therefore, w.p.1, for all large enough n, we have $BIC_n(\mathcal{M}_k) > BIC_n(\mathcal{M}_{k^*}) > c_k > 0$ for a some constant c_k. Therefore, we have that $\limsup_{n\to\infty} \hat{k}_n \leq k^*$ $w.p.1$. In contrast, if $\mathcal{M}_k \subset \mathcal{M}_{k*}$, then we have that:

$$\begin{aligned} BIC_n(\mathcal{M}_k) - BIC_n(\mathcal{M}_{k^*}) = n\Bigg\{ & \frac{1}{n}\ell_{\hat{\theta}^{k^*}_n}(y_{0:n-1}) - \frac{1}{n}\ell_{\hat{\theta}^{k}_n}(y_{0:n-1}) \\ & - \frac{(d_k - d_{k^*})\log n}{n}\Bigg\}, \\ & \to +\infty \, w.p.1, \end{aligned}$$

since $n \to \infty$, $\lim_{n\to\infty} \frac{1}{n}\ell_{\hat{\theta}^{k^*}_n}(y_{0:n-1}) - \frac{1}{n}\ell_{\hat{\theta}^{k}_n}(y_{0:n-1}) > 0$ by (1) of Proposition 33 and $\frac{(d_k - d_{k^*})\log n}{n} \to 0$ holds. Therefore $\liminf_{n\to\infty} \hat{k}_n \geq k^*$ w.p.1.

(2) Without loss of generality, we consider logarithm of evidence $\log m(y_{0;n-1})$. Then the claim follows directly from Proposition 31 and part (1) of this proposition. □

Therefore, BIC is strongly consistent for a general class of HMMs in the nested model setting we are considering here - with a model assumed to be a correctly specified one.

5.6.2 Asymptotic properties of AIC

We can be quite explicit about the behaviour of AIC. Making use of Proposition 30 gives rise to:

$$\begin{aligned} & \ell_{\hat{\theta}^{k}_n}(y_{0:n-1}) - \ell_{\hat{\theta}^{k^\star}_n}(y_{0:n-1}) \\ & \qquad = \frac{1}{2}\frac{\nabla^\top \ell_{\theta^k_\star}(y_{0:n-1})}{\sqrt{n}} \mathcal{J}^{-1}_{\theta^k_\star} \frac{\nabla \ell_{\theta^k_\star}(y_{0:n-1})}{\sqrt{n}} \\ & \qquad - \frac{1}{2}\frac{\nabla^\top \ell_{\theta^{k^\star}_\star}(y_{0:n-1})}{\sqrt{n}} \mathcal{J}^{-1}_{\theta^k_\star} \frac{\nabla \ell_{\theta^{k^\star}_\star}(y_{0:n-1})}{\sqrt{n}} + \epsilon, \end{aligned} \tag{5.40}$$

where $\epsilon = o(\log\log n)$, w.p1. Due to working with nested models, we have (immediately from the definition of $\mathcal{J}_{\theta^k_\star}$ and $\mathcal{J}_{\theta_\star}$):

$$\mathcal{J} := \mathcal{J}_{\theta^k_\star} = \begin{pmatrix} \mathcal{J}_{11} & \mathcal{J}_{12} \\ \mathcal{J}_{21} & \mathcal{J}_{22} \end{pmatrix} \in \mathbb{R}^{d_k \times d_k},$$

where $\mathcal{J}_{11} := \mathcal{J}_{\theta_\star}$, and $\mathcal{J}_{12}, \mathcal{J}_{21} = \mathcal{J}_{12}^\top$ are deduced from $\mathcal{J}_{\theta_\star}$. Similarly, $\nabla_\theta \ell_{\theta_\star}(y_{0:n-1})$ forms the upper $d_{k^\star}-$dimensional part of $\nabla \ell_{\theta^k_\star}(y_{0:n-1})$. We will use of the matrix equations implied by

$$\begin{gathered} \mathcal{J}\mathcal{J}^{-1} = I_{d_k} \iff \\ \begin{pmatrix} \mathcal{J}_{11} & \mathcal{J}_{12} \\ \mathcal{J}_{21} & \mathcal{J}_{22} \end{pmatrix} \begin{pmatrix} \mathcal{J}_{11}^{-1} & \mathcal{J}_{12}^{-1} \\ \mathcal{J}_{21}^{-1} & \mathcal{J}_{22}^{-1} \end{pmatrix} = \begin{pmatrix} I_{d_{k^\star}} & 0_{d_{k^\star}\times(d_k - d_{k^\star})} \\ 0_{(d_k - d_{k^\star})\times d_{k^\star}} & I_{(d_k - d_{k^\star})} \end{pmatrix}. \end{gathered}$$

Given the above nesting considerations, some cumbersome but otherwise straightforward calculations give:

$$\begin{aligned}&\frac{\nabla^\top \ell_{\theta_*^k}(y_{0:n-1})}{\sqrt{n}} J\left(\theta_\star^k\right)^{-1} \frac{\nabla \ell_{\theta_\star^k}(y_{0:n-1})}{\sqrt{n}} - \frac{\nabla^\top \ell_{\theta_\star^{k\star}}(y_{0:n-1})}{\sqrt{n}} J\left(\theta_*^{k^*}\right)^{-1} \frac{\nabla \ell_{\theta_\star^{k\star}}(y_{0:n-1})}{\sqrt{n}} \\ &\equiv \frac{\left\{M\nabla^\top \ell_{\theta_\star^k}(y_{0:n-1})\right\}^\top}{\sqrt{n}} D \frac{\left\{M\nabla^\top \ell_{\theta_\star^k}(y_{0:n-1})\right\}}{\sqrt{n}},\end{aligned} \tag{5.41}$$

where we have set:

$$\begin{aligned}M &:= \left((J^{-1})_{21}\left\{(J^{-1})_{22}\right\}^{-1}\right) \in \mathbb{R}^{(d_k - d_{k^\star}) \times d_{k^\star}}, \\ D &:= \left\{(J^{-1})_{22}\right\} \in \mathbb{R}^{(d_k - d_{k^\star}) \times (d_k - d_{k^\star})}.\end{aligned}$$

Consider the standard decomposition of the symmetric positive-definite D:

$$D = P\Lambda P^\top,$$

for orthonormal $P \in \mathbb{R}^{(d_k - d_{k^\star}) \times (d_k - d_{k^\star})}$ and diagonal $\Lambda \in \mathbb{R}^{(d_k - d_{k^\star}) \times (d_k - d_{k^\star})}$.

Assumption 13. *Define the martingale increments in* $\mathbb{R}^{d_k - d_{k^\star}}$, $k > k^\star$*:*

$$\tilde{h}_{\theta_\theta^k}(y_{-\infty:0}) := \left(\sqrt{\Lambda} P^\top M\right) h_{\theta_*^\star}(y_{-\infty:0}).$$

We have that $\mathbb{E}\left[\left(\tilde{h}_{\theta_\star^k}(y_{-\infty:0})\right)^2\right] > 0.$

Proposition 35. *Under Assumption 1 to Assumption 9 and Assumption 11 to Assumption 13, we have that, for* $k > k^*$,

$$\mathbb{P}\left(AIC_n\left(\mathcal{M}_k\right) < AIC_n\left(\mathcal{M}_{k^*}\right), infinitely\, often\, in\, n \geq 1\right) = 1.$$

Proof. Continuing from (5.40) and (5.41), the use of LIL for martingale increments will give that, w.p.1;

$$\begin{aligned}&\limsup_{n\to\infty} \frac{\sqrt{2}\left\{\ell_{\hat{\theta}_n^k}(y_{0:n-1}) - \ell_{\hat{\theta}_n^{k\star}}(y_{0:n-1})\right\}}{\log\log n} \\ &\geq \sup_{1 \leq j \leq d_k - d_{k^\star}} \mathbb{E}\left[\left(\tilde{h}_{\theta_\star^k}(y_{-\infty:0})\right)^2\right] > 0.\end{aligned}$$

As the difference $\ell_{\hat{\theta}_n^k}(y_{0:n-1}) - \ell_{\hat{\theta}_n^{k\star}}(y_{0:n-1})$ is size of $\Theta\left(\log\log n\right) - o\left(\log\log n\right)$ infinitely often, the result follows immediately. (The notation $\Theta(a_n)$ for a positive sequence $\{a_n\}$ means that the sequence of interest is upper and lower bounded by ca_n and $c^{'}a_n$ respectively for constants $0 < c < c^{'}$). □

This result implies that AIC might pick up a model that has more parameters than the true one has. In other words, as the sample size $n \to \infty$, with a positive probability, AIC will not necessarily select the correct index $k^\star$. If the additional assumption holds, then we can be more explicit, and prove

the following result for the behaviour of the AIC for increasing sample size. Indeed, we can show that AIC will overshoot in the sense that it will not select infinitely often the true model but a bigger one. Therefore, as $n \to \infty$, AIC will pick an over-fitted model infinitely often and is not a consistent Model Selection Criterion - in contrast with BIC, it is well known that AIC has desirable properties, e.g. with regards to prediction error (in many cases the model chosen by AIC attains the minimum maximum error in terms of prediction among models being considered), or its minimax optimality. Barron et al. (1999) is a comprehensive article on this topic and shows that minimax optimality of AIC holds in many cases, including the i.i.d., some non-linear models and for density estimation. For works on the efficiency of AIC terms of prediction see Shibata (1980, 1981); Shao (1997). AIC is equivalent to LOO cross-validation Stone (1977) for i.i.d.-type model structures. We refer to Ding et al. (2017, 2018) for a comprehensive review of AIC and BIC. Note that Yang (2005) shows that consistency of model selection and minimax optimality do not necessarily hold simultaneously. Our main focus in this work is asymptotic behaviour of criteria from a model selection viewpoint, so we will not further examine the prediction perspective. It also should be emphasised that even in the case of i.i.d. setting, the result such as Proposition 35 has yet to been proven (Claeskens and Hjort, 2008; Nishii, 1988).

5.6.3 A General Result

Following Sin and White (1996), one can generalise the above results for arbitrary penalty functions. Assume that we consider Information Criterion (IC) of the form:

$$IC_n\left(\mathcal{M}_k\right) = -\ell_{\hat{\theta}_n^k}(y_{0:n-1}) + pen_n(k), \tag{5.42}$$

for a penalty function $pen_n(k) \in \mathbb{R}$, (strictly) increasing in $k \geq 1$. For BIC, $pen_n(k) = d_k \log n$ and for AIC, $pen_n(k) = 2d_k$.

Proposition 36. *i) For $k' > k \geq 1$, if $pen_n(k)$ satisfies:*

$$\lim_{n\to\infty} \frac{pen_n(k') - pen_n(k)}{n} = 0,$$
$$\lim_{n\to\infty} \frac{pen_n(k') - pen_n(k)}{\log\log n} = +\infty,$$

under Assumption 1 to Assumption 9 and Assumption 12, the information criterion in (5.42) is strongly consistent.

ii) For $k' > k \geq 1$, if $pen_n(k)$ satisfies:

$$\lim_{n\to\infty} \frac{pen_n(k') - pen_n(k)}{n} = 0,$$
$$\lim_{n\to\infty} pen_n(k') - pen_n(k) = \infty,$$

then the information criterion in (5.42) is weakly consistent under underAssumption 1 to Assumption 9 and Assumption 12.

Proof. (1) It is an immediate generalisation of the proof of Proposition 34.

(2) Since $IC_n\left(\mathcal{M}_{\hat{k}}\right) \leq IC_n\left(\mathcal{M}_k\right)$ always holds for large enough n by the assumption, it suffices to show that for any $\epsilon > 0$ and fixed model k,

$$\lim_{n\to\infty} \mathbb{P}\left(IC_n\left(\mathcal{M}_k\right) - IC_n\left(\mathcal{M}_{k^\star}\right) > \epsilon\right) = 1.$$

First, let us consider the case when $\mathcal{M}_k \subset \mathcal{M}_{k^\star}$. Then, for any $\epsilon > 0$, we have that:

$$\begin{aligned} &\mathbb{P}\left(IC_n\left(\mathcal{M}_k\right) - IC_n\left(\mathcal{M}_{k^\star}\right) > \epsilon\right) \\ &= \mathbb{P}\left(\ell_{\hat{\theta}_n^{k^\star}}(y_{0:n-1}) - \ell_{\hat{\theta}_n^{k}}(y_{0:n-1}) \geq \left(pen_n(k) > pen_n(k^\star)\right)\right) \to 1. \end{aligned}$$

The limit follows from (1) of Proposition 33, as the random variable on the left side of the inequality above diverges to $+\infty$ w.p1. and this convergence implies convergence in probability. This result implies directly $\lim_{n\to\infty} \mathbb{P}\left(\hat{k}_n \geq k^\star\right) = 1$. Next we consider the case where $\mathcal{M}_{k^\star} \subset \mathcal{M}_k$. We obtain:

$$\mathbb{P}\left(IC_n\left(\mathcal{M}_k\right) \leq IC_n\left(\mathcal{M}_{k^*}\right)\right) \leq \mathbb{P}\left(2\left\{\ell^k_{\hat{\theta}^k_n}(y_{0:n-1}) - \ell^{k^\star}_{\hat{\theta}^{k^\star}_n}(y_{0:n-1})\right\} \geq \left(pen_n(k) > pen_n(k^\star)\right)\right). \quad (5.43)$$

As we have already shown, $2\left\{\ell^k_{\hat{\theta}^k_n}(y_{0:n-1}) - \ell^{k^\star}_{\hat{\theta}^{k^\star}_n}(y_{0:n-1})\right\} \Rightarrow Z^\top \mathcal{J}_{\theta^k_\star}{}^{-1} Z - Z^\top_{1:d_{k^\star}} \mathcal{J}_{\theta^k_\star}{}^{-1} Z_{1:d_{k^\star}} =: Z_0$. Continuing from (5.43), since $\mid Z_0 \mid < \infty$, w.p1. for any $\epsilon > 0$ we can have some n_0 so that for all $n_1 \geq n_0$, $\mathbb{P}\left(Z_0 \geq \left(pen_n(k) > pen_n(k^\star)\right)\right) < \epsilon$. Thus, for all n large enough, we have that:

$$\begin{aligned} &\mathbb{P}\left(IC_n\left(\mathcal{M}_k\right) \leq IC_n\left(\mathcal{M}_{k^*}\right)\right) \\ &\leq \mathbb{P}\left(2\left\{\ell^k_{\hat{\theta}^k_n}(y_{0:n-1}) - \ell^{k^\star}_{\hat{\theta}^{k^\star}_n}(y_{0:n-1})\right\} \geq \left(pen_{n_1}(k) > pen_{1}(k^\star)\right)\right) \\ &\to \mathbb{P}\left(Z_0 \geq \left(pen_n(k) > pen_n(k^\star)\right)\right) < \epsilon. \end{aligned}$$

Thus, we conclude that $\lim_{n\to\infty} \mathbb{P}\left(IC_n\left(\mathcal{M}_k\right) \leq IC_n\left(\mathcal{M}_{k^*}\right)\right) = 0$ and the claim follows. □

The above results indicate that to have strong or weak consistency, a penalty function should depend on the sample size and grow to infinity at a specific rate. Thus the criterion whose penalty function does not depend on the sample size will not be necessarily consistent in terms of both strong and weak sense. Besides, $\log n$ is not the slowest rate which attains strong consistency almost surely. This observation implies that another criterion which has a slower penalty function might be strongly consistent. For instance, Hannan and Quinn (1979) proposes the criterion whose penalty function is given by $-2c \log\log d$ where c is a constant such that $c > 1$. Although we do not study this criterion in this article, it might possess strong consistency for HMMs.

5.7 Particle approximation of AIC and BIC

BIC and AIC can be used for model selection for HMMs but are typically impossible to calculate analytically due to the intractability of the likelihood function for general HMMs. Thus, an approximation technique is required. We adopt the computational approach of Poyiadjis et al. (2011) which,

for completeness, we briefly review in this Section. See also subsubsection 4.5.1. It involves a particle filtering algorithm coupled with a recursive construction for an integral approximation.

The description follows closely Poyiadjis et al. (2011). The marginal Fisher identity gives,

$$\nabla \ell_\theta(y_{0:n}) = \int_{\mathsf{X}} \nabla \log p_\theta(x_n, y_{0:n}) p_\theta(x_n | y_{0:n}) \mu(dx_n).$$

At step n, let $(x_n^{(i)}, W_n^{(i)})_{i=1}^N$ be a particle approximation of $p_\theta(dx_n \mid y_{1:n})$, with standardised weights, i.e. $\sum_i W_n^{(i)} = 1$, obtained via some particle filtering algorithm (see Doucet and Johansen (2009) for instance), so that,

$$\nabla \ell_\theta(y_{0:n-1}) \simeq \sum_{i=1}^N W_n^{(i)} \nabla \log p_\theta(x_n^{(i)}, y_{1:n}). \tag{5.44}$$

We explore the unknown quantity $\nabla_\theta \log p_\theta(x_n, y_{0:n})$. First, observe that:

$$p_\theta(x_n, y_{0:n}) = g_\theta(y_n \mid x_n) p_\theta(y_{0:n-1}) \int_{\mathsf{X}} q_\theta(x_n \mid x_{n-1}) p_\theta(x_{n-1} \mid y_{0:n-1}) \mu(dx_{n-1}). \tag{5.45}$$

This implies that:

$$\begin{aligned} \nabla p_\theta(x_n, y_{0:n}) = p_\theta(y_{1:n-1}) g_\theta(y_n \mid x_n) \int_{\mathsf{X}} q_\theta(x_n \mid x_{n-1}) p_\theta(x_{n-1} \mid y_{0:n-1}) \times \\ \{\nabla \log g_\theta(y_n \mid x_n) + \nabla \log q_\theta(x_n \mid x_{n-1}) + \nabla p_\theta(x_{n-1},\, y_{0:n-1})\}\, \mu(dx_{n-1}). \end{aligned} \tag{5.46}$$

At step $n-1$, let $(x_{n-1}^{(i)}, w_{n-1}^{(i)})_{i=1}^N$ be a particle approximation of the filtering distribution $p_\theta(dx_{n-1} \mid y_{1:n-1})$ and $\left\{\alpha_{n-1}^{(i)}\right\}_{i=1}^N$ be a sequence of approximations to $\left(\nabla \log p_\theta(x_{n-1}^{(i)}, y_{0:n})\right)_{i=1}^N$. Equations (5.45), (5.46) suggest the following recursive approximation of $\nabla_\theta \log p_\theta(x_n^{(i)}, y_{0:n})$, for $1 \le i \le N$,

$$\begin{aligned} \alpha_n^{(i)} = & \frac{\sum_{j=1}^N W_{n-1}^{(j)} q_\theta(x_n^{(i)} \mid x_{n-1}^{(j)})}{\sum_{k=1}^N W_{n-1}^{(k)} q_\theta(x_n^{(i)} \mid x_{n-1}^{(k)})} \\ & \times \left\{\nabla \log g_\theta(y_n \mid x_n^{(i)}) + \nabla \log q_\theta(x_n^{(i)} \mid x_{n-1}^{(j)}) + \alpha_{n-1}^{(j}\right\}. \end{aligned} \tag{5.47}$$

Thus, from (5.44), one obtains an estimate of the score function at step n, as:

$$\nabla \ell_\theta(y_{0:n-1}) \simeq \sum_{i=1}^N W_n^{(i)} \alpha_n^{(i)}. \tag{5.48}$$

The calculation in (5.47), and the adjoining particle filtering algorithm, can be applied recursively to provide the approximation of the score function in (5.48) for $n = 0, 1, \cdots$. Note that the computational cost is $\mathcal{O}(N^2)$, but is robust for increasing n as it is based on the approximation of the filtering distributions rather than the smoothing ones, see results and comments on this point in Poyiadjis et al. (2011); Del Moral et al. (2015).

Moreover, Poyiadjis et al. (2011) uses the score function estimation methodology to propose an

online gradient ascent algorithm for obtaining an MLE-type parameter estimate. In more detail, note that $\nabla \log p_\theta(y_n|y_{0:n-1}) = \nabla \ell_\theta(y_{0:n}) - \nabla \ell_\theta(y_{0:n-1})$. Then one obtains the following recursion:

$$\begin{aligned}\theta_{n+1} &= \theta_n + \gamma_{n+1} \log p_\theta(y_n \mid y_{0:n-1})_{\theta=\theta_n} \\ &= \gamma_{n+1} \int_{\mathsf{X}} \nabla_\theta \log p_\theta(x_n, y_{0:n})|_{\theta=\theta_n} p_\theta(x_n \mid y_{0:n})_{\theta=\theta_n} \mu(dx_n) \\ &- \gamma_{n+1} \int_{\mathsf{X}} \nabla_\theta \log p_\theta(x_{n-1}, y_{0:n-1})|_{\theta=\theta_{n-1}} p_\theta(x_{n-1}|y_{0:n-1})_{\theta=\theta_{n-1}} \mu(dx_{n-1}),\end{aligned} \tag{5.49}$$

where $\{\gamma_k\}_{k\geq 1}$ is a real-valued decreasing sequence with:

$$\sum_{k=1}^{\infty} \gamma_k = \infty, \sum_{k=1}^{\infty} \gamma_k^2 < \infty.$$

To deduce an online algorithm, following ideas in Le Gland and Mevel (1997), intermediate quantities involved in the recursions in (5.44)-(5.47) are calculated at different, consecutive parameter values. See Poyiadjis et al. (2011) for more details, and Le Gland and Mevel (1997); Tadic (2010) for analytical studies on the convergence properties of the algorithm. In particular, under strict conditions, and cases or trivial models, the algorithm is shown to converge to the maximiser θ_* of the limiting function of $\theta \mapsto \ell_n(\theta)/n$, as $n \to \infty$.

Remark 14. *In our setting, we want to use numerical studies to illustrate the theoretical results obtained for AIC and BIC, so we will use the outcome of the online recursion as a proxy for the MLE. Then, the AIC and BIC will be approximated by running a particle filter for the chosen MLE value to obtain an approximation of the log-likelihood of the data at this parameter value.*

5.8 Empirical Study

Motivated by the numerics in Pitt et al. (2014), we consider the following stochastic volatility model (labelled as $\mathcal{SV}$):

$$\mathcal{SV}: \begin{cases} X_t = \phi X_{t-1} + W_t, \\ Y_t = exp(X_t/2)V_t, \ t \geq 0 \end{cases}$$

and the one with jumps (labelled as $\mathcal{SVJ}$):

$$\mathcal{SVJ}: \begin{cases} X_t = \phi X_{t-1} + W_t, \\ Y_t = exp(X_t/2)V_t + q_t J_t, \ t \geq 0 \end{cases}$$

where $W_t \sim \mathcal{N}(0, \sigma_X^2)$, $V_t \sim \mathcal{N}(0,1)$, $J_t \sim \mathcal{N}(0, \sigma_J^2)$ and $q_t \sim Bernoulli(p)$, all variables assumed independent over the time index $t \geq 1$. In both cases, $X_0 = 0$. Here $Bernoulli(p)$ stands for the Bernoulli distribution with parameter p. **??** shows two sets of 10^4 simulated observations, one from

$\mathcal{SV}$ and one from $\mathcal{SVJ}$, under the the corresponding true parameter values:

$$(\phi, \sigma_X) = (0.9, \sqrt{0.3}),$$
$$(\phi, \sigma_X, \sigma_J, p) = (0.9, \sqrt{0.3}, \sqrt{0.6}, 0.6).$$

These simulated data will be used in the experiments that follow. Scenario 1 (resp. Scenario 2) corresponds to the case when the true model is $\mathcal{SVJ}$ (resp. $\mathcal{SV}$). We will compare the two classes of models, using AIC and BIC, in both Scenarios. The estimated parameter values for $\mathcal{SV}$ and $\mathcal{SVJ}$, and then estimates for AIC and BIC using a particle filter, via the method of Poyiadjis et al. (2011), reviewed in subsection 5.7. Note that, as we have established in this work, BIC is expected to be consistent for both Scenarios, whereas AIC only for the first Scenario.

Figure 8 shows estimated parameter values for $\mathcal{SV}$, $\mathcal{SVJ}$, for both simulated sample size, sequentially in the data size, using the online version of the method of Section 6, with $N = 200$ particles. (We also tried a larger number of particles, with similar results.) To further investigate the stability of the algorithm we summarise in Figure 9 and Figure 10 estimates of AIC and BIC for the two models from $R = 200$ replications of the same algorithm, for different sample sizes. Figure 9 corresponds to Scenario 1 and Figure 10 to Scenario 2. All results obtained seem to indicate that the numerical algorithm used for approximating AIC and BIC is fairly robust in all cases. Also, it appears that in the challenging Scenario 2, even with$n = 10^4$ observations, the box plots do not seem to provide any strong evidence in favour of true model $\mathcal{SV}$.

Table 2 shows results from the same $R = 200$ replications for the estimation of AIC and BIC for each of the two models and two simulated datasets. In agreement with our theory, BIC appears more robust (than AIC) at choosing the correct model for the dataset simulated from $\mathcal{SV}$. Figure 11 plots differences in AIC and BIC in Scenario 2, sequentially in the sample size, more accurately, a proxy of the differences, see Remark 14. To be precise, the blue line denotes the 'path' of AIC($\mathcal{SV}$)−AIC($\mathcal{SVJ}$), and the red line denotes the one of BIC($\mathcal{SV}$)−BIC($\mathcal{SVJ}$). Since model $\mathcal{SV}$ is true in this case, the difference should be lower than zero for large enough n if the used IC were consistent. As one can see, the difference in BIC is always negative after a large enough sample size n. In contrast, and agreement with our theory, the difference in AIC never has such property. For instance, sometime after $n = 10^4$, the difference increased and exceeded the zero line. This is a clear empirical manifestation of Proposition 35; so, whereas in the previous plots the deficiency of AIC was difficult to showcase when looking at *fixed* sample sizes, such deficiency became clear when we look at the evolution of AIC as a function of sample size.

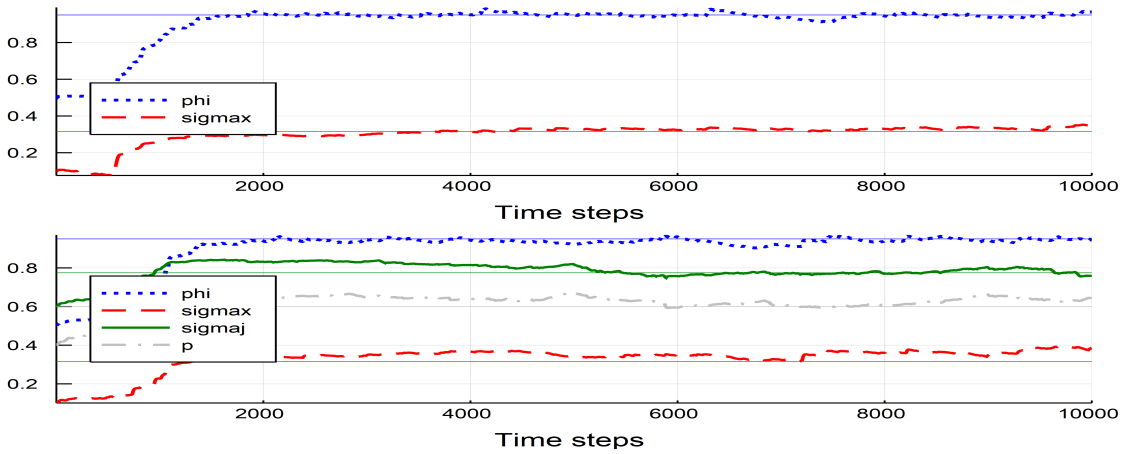


Figure 8: Estimated parameters for the $\mathcal{SV}$(to panel) and $\mathcal{SVJ}$(bottom panel) models as obtained - sequentially in time - via the data simulated from the $\mathcal{SV}$(top panel) and $\mathcal{SVJ}$(bottom panel) models respectively and the algorithm reviewed in Section 6 with $N = 200$ particles. The horizontal lines indicate the true parameter values in each case.

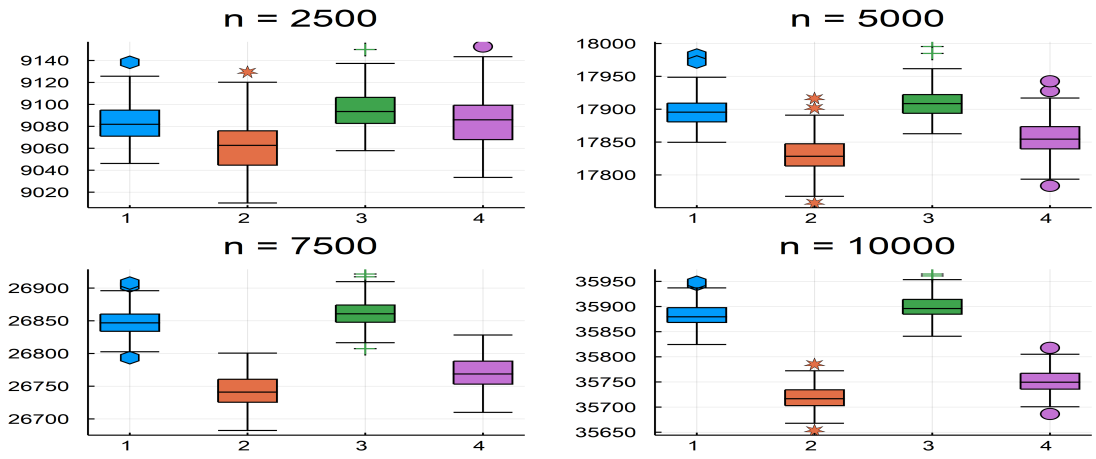


Figure 9: Boxplots for Scenario 1 (SVJ is true) from $R = 200$ estimates of AIC and BIC and various observation sizes. Blue: AIC($\mathcal{SV}$), Orange: AIC($\mathcal{SVJ}$), Green: BIC($\mathcal{SV}$), Purple: BIC($\mathcal{SVJ}$).

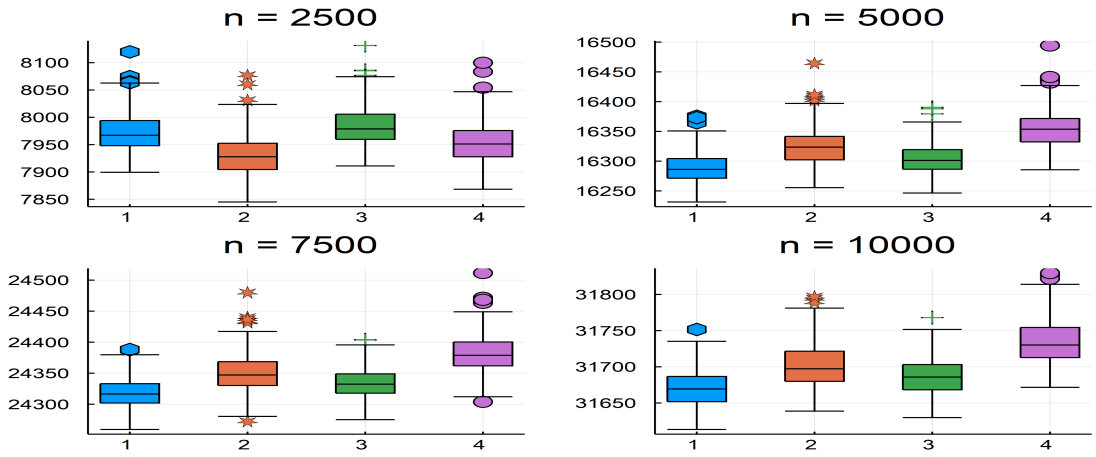


Figure 10: Boxplots for Scenario 2 (SV is true) from $R = 200$ estimates of IC and various observation sizes. Blue: AIC($\mathcal{SV}$), Orange: AIC($\mathcal{SVJ}$), Green: BIC($\mathcal{SV}$), Purple: BIC($\mathcal{SVJ}$).

n	$2,500$	$5,000$	$7,500$	$10,000$	n	$2,500$	$5,000$	$7,500$	$10,000$
AIC($\mathcal{SV}$)	$\frac{32}{200}$	$\frac{4}{200}$	$\frac{0}{200}$	$\frac{0}{200}$	AIC($\mathcal{SV}$)	$\frac{154}{200}$	$\frac{161}{200}$	$\frac{153}{200}$	$\frac{158}{200}$
AIC($\mathcal{SV}J$)	$\frac{168}{200}$	$\frac{196}{200}$	$\frac{200}{200}$	$\frac{200}{200}$	AIC($\mathcal{SVJ}$)	$\frac{46}{200}$	$\frac{39}{200}$	$\frac{37}{200}$	$\frac{42}{200}$
BIC($\mathcal{SV}$)	$\frac{51}{200}$	$\frac{5}{200}$	$\frac{0}{200}$	$\frac{0}{200}$	BIC($\mathcal{SV}$)	$\frac{179}{200}$	$\frac{173}{200}$	$\frac{184}{200}$	$\frac{192}{200}$
BIC($\mathcal{SVJ}$)	$\frac{149}{200}$	$\frac{195}{200}$	$\frac{200}{200}$	$\frac{200}{200}$	BIC($\mathcal{SVJ}$)	$\frac{25}{200}$	$\frac{27}{200}$	$\frac{16}{200}$	$\frac{8}{200}$

Table 2: Results after $R = 200$ replications of the approximation algorithm with $N = 200$ particles. The 1st (resp. 2nd) row in the table shows the fraction of the replications where the estimated AIC is smaller for the $\mathcal{SV}$ model (resp. $\mathcal{SVJ}$ model) for different number of sample sizes; rows 3 and 4 show similar results for BIC. The left (resp. right) side of the table shows results for the data simulated from $\mathcal{SVJ}$ (resp. $\mathcal{SV}$).

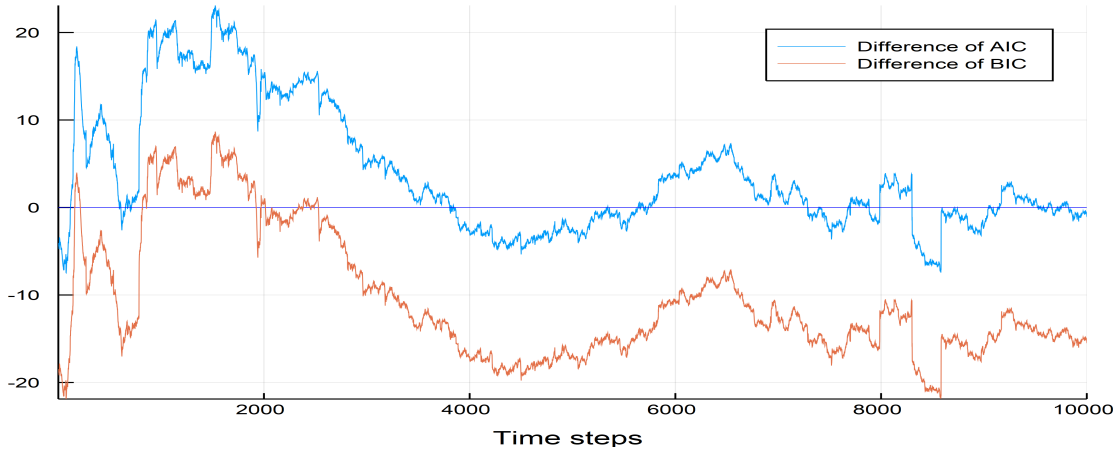


Figure 11: The path of differences in AIC and BIC in Scenario 2 ($\mathcal{SV}$ is the true model). That is, the blue line show the approximated value of AIC($\mathcal{SV}$)−AIC($\mathcal{SVJ}$) as a function of data size, and the red line the corresponding function for BIC($\mathcal{SV}$)−BIC($\mathcal{SVJ}$).

5.9 Conclusion and remarks

We have investigated the asymptotic behaviour of BIC, the evidence and AIC for nested HMMs, and have derived new results concerning their consistency properties. Our work shows that BIC – and the evidence – are strongly consistent for a general class of HMMs. In contrast, for a similarly posed Model Selection problem, AIC is not even weakly consistent. Our study focuses on asymptotics for increasing data size, so we do not investigate finite sample-size results for BIC, evidence and AIC, such as optimality properties. It is well-known that AIC is minimax-rate optimal but BIC is not in many cases, see e.g. Barron et al. (1999). We conjecture this might also be the case for general HMMs.

The technique of constructing stationary, ergodic processes by introducing a backward infinite extension of the observations (see subsection 5.4) has been used in many other studies, even beyond HMMs. For instance, Douc et al. (2020) use this approach to study posterior consistency for a class of partially observed Markov models, Lehéricy (2018) uses it to investigate non-asymptotic behaviour of MLE for (finite state space) HMMs, Le Corff et al. (2013) apply the technique within an online EM setting for HMMs in a no-model-correctness setting, Diel et al. (2020) consider more general classes of latent variable models.

We note that asymptotic results about the MLE for HMMs have recently been obtained under weaker conditions. See, e.g., Douc et al. (2011b, 2016b) for developments that go beyond compact spaces. Here, we have worked with strict assumptions on the state space (see Assumption 3), so that we obtain an important first set of illustrative results for Model Selection Criteria, avoiding at the same time an overload of technicalities. Future investigations are expected to further weaken the conditions we have used here.

Our results are obtained in the context of nested models, where a model is assumed to be the true data-generating one. There are challenges when trying to move beyond the Model-Correctness setting. As we have described in the first parts of the paper, Douc and Moulines (2012) show that the MLE converges a.s. even for misspecified models under mild assumptions. However, a CLT for the MLE

in the context of general state-space misspecified HMMs has yet to been proven.To the best of our knowledge, only Pouzo et al. (2016) obtain such a result for a finite state-space X. Thus, extending our results to non-nested settings or/and ones where one does not assume correctness of a model, is a non-trivial undertaking that requires extensive further research. Also, we note that AIC is asymptotically prediction efficient in some misspecified models whilst BIC is not. The above discussion suggests that investigating asymptotic behaviour of model selection criteria under No-Model-Correctness for general HMM models is an important open problem that requires further research.

One can use alternative numerical algorithms instead of the one we have used here, and describe in subsection 5.8, see e.g. the approach in Olsson and Alenlöv (2020). Note that the numerical algorithm used in the paper is mostly a tool for illustrating our theoretical findings, which is the main focus of our work. The numerical study shown in the paper already delivers the points stemming from the theory, so we have refrained from describing/implementing further methods to avoid diverting attention from our main findings

From a practitioner point of view, our results and numerical study indicate that AIC can wrongly select the more complex model due to ineffective penalty term. Critically, this can be difficult to assess using standard experiments. Our study has shown that one needs to investigate the evolution of AIC against data size to clearly highlight its deficiency in the context of a numerical study. We stress here that in the numerical experiment we have knowingly used models for which several of the stated Assumptions will not hold (maybe most notably, the strong mixing Assumption 3). The aim is to illustrate at least numerically, that while our assumptions are standard in the literature, they serve for simplifying the path to otherwise too technical derivations and provide results that are expected to hold in much more general settings.

6 Online Smoothing for Diffusion Processes Observed with Noise

6.1 Introduction

We introduce a methodology for online estimation of smoothing expectations for a class of additive functionals, in the context of a rich family of diffusion processes (that may include jumps), observed at discrete-time instances. We overcome the unavailability of the transition density of the underlying SDE by working on the augmented pathspace. The new method can be applied, for instance, to carry out online parameter inference for the designated class of models. Algorithms defined on the infinite-dimensional pathspace have been developed in the last years mainly in the context of MCMC techniques. The main benefit is the achievement of mesh-free mixing times for the practical time-discretised algorithm used on a PC. Our own methodology sets up the framework for infinite-dimensional online filtering, an important positive practical consequence is the construct of estimates with variance that does not increase with decreasing mesh-size. Besides regularity conditions, our method is, in principle, applicable under the weak assumption, relatively to restrictive conditions often required in the MCMC or filtering literature of methods defined on pathspace -- that the SDE covariance matrix is invertible.

Research in Hidden Markov Models (HMMs) has, thus far, provided effective online algorithms for the estimation of expectations of the smoothing distribution for the case of a class of additive functionals of the underlying signal. Such methods necessitate knowledge of the transition density of the Markovian part of the model between observation times. We carry out a related exploration for the (common in applications) case when the signal corresponds to a diffusion process, thus we are faced with the challenge that such transition densities are typically unavailable. Standard data augmentation schemes that work with the multivariate density of a large enough number of imputed points of the continuous-time signal will lead to ineffective algorithms. The latter will have the abnormal characteristic that, for given Monte-Carlo iterates, the variability of the produced estimates will increase rapidly as the resolution of the imputation becomes finer. One of the ideas underpinning the work in this paper is that development of effective algorithms instead requires respecting the structural properties of the diffusion process, thus we build up imputation schemes on the infinite-dimensional diffusion pathspace itself. As a consequence, the time-discretised algorithm used in practice on a PC will be stable under mesh-refinement.

We consider continuous-time jump-diffusion models observed at discrete-time instances. The d_x-dimensional process, $X = \{X_t; t \geq 0\}$, $d_x \geq 1$ is defined via the following time-homogeneous stochastic differential equation (SDE), with $X_{t-} := \lim_{s \uparrow t} X_t$:

$$dX_t = b_\zeta(X_{t-})dt + \sigma_\zeta(X_{t-})dW_t + dJ_t, \tag{6.1}$$

with $X_0 \sim \pi(dx)$, $t \geq 0$. The solution X is driven by the d_w-dimensional Brownian motion, $\{W_t; t \geq 0\}$, $d_w \geq 1$, the compound Poisson process, $\{J_t; t \geq 0\}$ and the initial distribution π_0. The SDE involves a drift function $b_\zeta : \mathbb{R}^{d_x} \mapsto \mathbb{R}$ and coefficient matrix $\sigma_\zeta : \mathbb{R}^{d_x} \mapsto \mathbb{R}^{d_x \times d_w}$ with parameter $\zeta \in \mathbb{R}^p$. Let $\{N_t; t \geq 0\}$ be a Poisson process with intensity function $\lambda_\nu(\cdot)$, and $\{\xi_k\}_{k \geq 1}$ an i.i.d. sequence of random variables each with pdf $h_\mu(\cdot)$; the cadlag process J is determined as $J_t = \sum_{i=1}^{N_t} \xi_i$

for parameters $\eta := (\nu, \mu)$, $\nu \in \mathbb{R}^q$, $\mu \in \mathbb{R}^r$ $q, r \geq 1$. We set:

$$\theta := (\zeta, \eta). \tag{6.2}$$

We work under standard assumptions (e.g. linear growth, Lipschitz continuity for b_ζ, σ_ζ) that guarantee a unique global solution of (6.1), in a weak sense, see e.g. Øksendal and Sulem (2007).

SDE (6.1) is observed with noise at discrete-time instances $0 = t_0 < t_1 < t_2 < \cdots < t_n$, $n \geq 1$. Without loss of generality, we assume equidistant observation times, with $\Delta := t_1 - t_0$.We consider data $Y_{t_0}, \cdots Y_{t_n}$ and for simplicity we set:

$$\begin{aligned} x_i &:= X_{t_i},\ y_i := Y_{t_i}, \\ \mathcal{F}_0 &= \sigma(X_0),\ \mathcal{F}_i := \sigma\left(\{X_s; s \in [t_{i-1}, t_i]\}\right), \end{aligned}$$

for $0 \leq i \leq n$. We also assume:

$$\left[Y_{t_i} \mid \{Y_{t_j}; j < i\}, X_s; s \in [0, t_i]\right] \sim g_\theta(dY_{t_i} \mid Y_{t_{i-1}}, \mathcal{F}_i), \tag{6.3}$$

for conditional distribution $g_\theta(\cdot \mid Y_{t_{i-1}}, \mathcal{F}_i)$ on $\mathbb{R}^{d_y}$, $d_y \geq 1$ under the convention $Y_{t_{-1}} = y_{-1} = \emptyset$ for $0 \leq i \leq n$. We write:

$$[x_i \mid x_{i-1}] \sim f_\theta(dx_i \mid x_{i-1}), \tag{6.4}$$

where $f_\theta(\cdot \mid x_{i-1})$ is the transition distribution of the driving SDE process (6.1). We assume existence of density functions for $g_\theta(dy_i \mid y_{i-1}, \mathcal{F}_i)$and $f_\theta(dx_i \mid x_{i-1})$, and, with some abuse of notation, we write $g_\theta(dy_i \mid y_{i-1}, \mathcal{F}_i) = g_\theta(y_i \mid y_{i-1}, \mathcal{F}_i)$ and $f_\theta(dx_i \mid x_{i-1}) = f_\theta(x_i \mid x_{i-1})$ where dy_i, dx_i denote the relevant Lebesgue measures.

Our work is relevant under the following regime.

Assumption 14. *For any $0 \leq i \leq n$, the transition density $f_\theta(x_i \mid x_{i-1})$ is intractable and the likelihood density $g_\theta(y_i \mid y_{i-1}, \mathcal{F}_i)$ is analytically available.*

The intractability of the transition density $f_\theta(x' \mid x)$ will pose challenges in the main inferential problems that this paper aims to address.

Models defined via (6.1)-(6.4) are extensively used, e.g., in finance and econometrics, for instance for capturing the market microstructure noise, see Aït-Sahalia et al. (2005); Hansen and Lunde (2006). The above setting belongs in the general class of hidden Markov models (HHMs), with a signal defined in continuous-time. See Cappé et al. (2005); Douc et al. (2014) and subsection 4.3 for a general treatment of HMMs fully specified in discrete-time.

A number of methods have been suggested in the literature for approximating the transition density – mainly in the case of no-jump component – including: asymptotic expansion techniques (Ait-Sahalia and Yu, 2008; Kessler, 1997; Aït-Sahalia et al., 2005; Aït-Sahalia, 2002); martingale estimating functions Kessler and Sørensen (1999); generalized method of moments (Hansen and Scheinkman, 1993); Monte-Carlo approaches (Wagner, 1989; Durham, 2003; Beskos et al., 2006). See, e.g., Kessler et al. (2012) and references therein for a detailed review.

For a given sequence $\{a_m\}_{m\geq 0}$, we use the notation $a_{i:j} := (a_i, \cdots a_j)$, for integers $i \leq h$. Let $p_\theta(y_{0:n})$ denote the joint density of $y_{0:n}$. Throughout the paper, $p_\theta(\cdot)$ is used generically to represent probability distributions or densities of random variables appearing as arguments in $p_\theta(\cdot)$. Consider the maximum likelihood estimator (MLE):

$$\hat{\theta}_n := \arg\max_{\theta\in\Theta} \log p_\theta(y_{0:n}). \tag{6.5}$$

Except for limited cases, one cannot obtain the MLE analytically for HMMs (even for discrete-time signal) due to the intractability of $p_\theta(y_{0:n})$.We have set up the modelling context for this work. The main contributions of this section in this setting – several of which relate with overcoming the intractability of the transition density of the SDE – will be as follows:

i) We present an *online* algorithm that delivers Monte-Carlo estimators of smoothing expectations:

$$\mathcal{S}_{\theta,n} := \int S_\theta(\mathbf{x}_{0:n}) p_\theta(d\mathbf{x}_{0:n} \mid y_{0:n}),\ n \geq 1, \tag{6.6}$$

on the class of additive functionals $S_\theta(\cdot)$ of the structure:

$$S_\theta(\mathbf{x}_{0:n}) := \sum_{k=0}^{n} s_{\theta,k}(x_{k-1}, \mathbf{x}_k), \tag{6.7}$$

(under the conventions $x_{-1} = \emptyset$, $\mathbf{x}_0 = x_0$). This is seen as the means to solve some particular problem. The definition of $\{\mathbf{x}_k\}_{k\geq 0}$ will be provided in the main text; for now we stress that $\mathbf{x}_k$ is a pathspace-valued element that corresponds to an 1–1transform of $\{X_s; s \in [t_{i-1}, t_i]\}$.

ii) We take advantage of the new approach to show numerical applications, with emphasis on carrying out *online* parameter inference for the designated class of models via a gradient-ascent approach (in a Robbins-Monro stochastic gradient framework). A critical aspect of this particular online algorithm (partly likelihood based, when concerned with parameter estimation; partly Bayesian, with regards to identification of filtering/smoothing expectations) is that it delivers estimates of the evolving score function, of the model parameters, together with particle representations of the filtering distributions, through a *single passage* of the data. This is a unique favourable algorithmic characteristic, when constrasted with alternative algorithms with similar objectives, such as, e.g., Particle MCMC Andrieu et al. (2010), or SMC^2 Chopin et al. (2013).

iii) In this work, we will not characterise analytically the size of the time-discretisation bias relevant to the SDE models at hand, and are content that: (1) the bias can be decreased by increasing the resolution of the numerical scheme (typically an Eyler-Maruyama (1.48) one, or some other Taylor scheme, see e.g. Kloeden and Platen (2013)); (2) critically, the algorithms are developed in a manner that their performance is stable when increasing the resolution of the time-discretisation method, as the reader will notice that the algorithms are (purposely) first defined on the infinite-dimensional pathspace, and SDE paths are only discretised when implementing the algorithm on a PC (to allow for finite computations).

iv) Our method draws inspiration from earlier works, in the context of online filtering for discrete-

time HMMs and infinite-dimensional pathspace MCMC methods. The complete construct is novel; one consequence of this is that it is applicable, in principle, for a wide class of SDEs, under the following weak assumption (relatively to restrictive conditions often imposed in the literature of infinite-dimensional MCMC methods).

Assumption 15. *The diffusion covariance matrix function:*

$$\Sigma_\zeta(x) := \sigma_\zeta(x)\sigma_\zeta^\top(x) \in \mathbb{R}^{d_x \times d_x} \tag{6.8}$$

is invertible, for all relevant x, ζ.

Thus, the methodology does not apply as demonstrated here only for the class of hypoelliptic SDEs.

An elegant solution to the problem posed above, for the case of a standard HMM with discrete-time signal of known transition density $f_\theta(x' \mid x)$, is given in Del Moral et al. (2010); Poyiadjis et al. (2011) as we studied. Our own work overcomes the unavailability of the transition density in the continuous-time scenario by following closely Poyiadjis et al. (2011) but augmenting the hidden state with the complete continuous-time SDE path. Related augmentation approaches in this setting – though for different inferential objectives – have appeared in Fearnhead et al. (2008); Ströjby and Olsson (2009); Gloaguen et al. (2018) where the augmented variables are derived via the Poisson estimator of transition densities for SDEs (without jumps), introduced in Beskos et al. (2006), and in Särkkä and Sottinen (2008) where the augmentation involves indeed the continuous-time path (the objective therein is to solve the filtering problem and the method is applicable for SDEs without jumps and additive Wiener noise).

A Motivationg Example. Figure 12 shows estimates of the score function, evaluated at the true parameter value $\theta^\star$, for parameters θ_3 of the Ornstein Uhlenbeck (OU) process, $dX_t = \theta_1(\theta_2 - X_t)dt + \theta_3 dW_t$, $X_0 = 0.0$, for $n = 10$ observations $y_i = x_i + \epsilon_i$, $\epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, 0.1^2)$. Data were simulated from $\theta^\star(0.5, 0.0, 0.4)$ with an Euler-Maruyama scheme of $M = 10$ grid points per unit of time. Figure 12 illustrates the 'abnormal' effect of a standard data-augmentation scheme, where for $N = 100$ particles, the Monte-Carlo method (see later sections for details) produces estimates of increasing variability as algorithmic resolution increases with $M = 10, 50, 150, 200$, i.e., as it approaches the 'true' resolution used for the data generation.

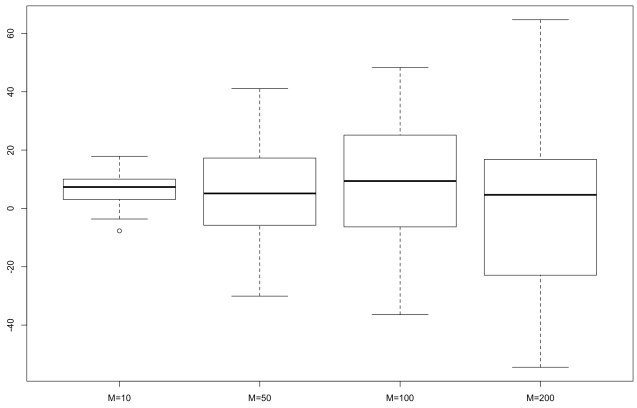


Figure 12: Boxplots of estimated score functions of θ_3 for OU process over $R = 50$ experiment replications. $N = 100$ particles were used in all cases, for the same $n = 10$ data-points.

The rest of the paper is organised as follows. In subsection 6.2 we give basics of inference for discretely observed diffusions. Then we briefly review the forward only particle smoothing method when the transition density $f_\theta(x' \mid x)$ is tractable in subsection 6.3. Then subsection 6.4 is devoted to providing our approach to overcome the intractability of the transition density $f_\theta(x' \mid x)$. We first develop a technique for SDEs without jumps, then we generalise the result to SDEs with jumps. Our methods are based on introducing an appropriate bridge process. Using such technique, we formulate the problem on on diffusion pathspace, then develop forward only particle smoothing for discretely observed SDEs in subsection 6.5. By merging forward only particle smoothing and the Robbins–Monro, we develop an online gradient-ascent for partially observed SDEs in subsection 6.6. We use the algorithm to present some numerical results to quantify its numerical stability and performance in subsection 6.7, then conclude in subsection 6.8

6.2 Basics of inference for discretely observed diffusions

When it comes to parameter inference for discretely observed diffusions, the common problem for both frequentist and Bayesian approach is intractability of the law of $[X_t; 0 \leq t \leq T \mid X_0 = x, X_T = x']$. We review several attempts to overcome this problem, in particular, we focus on the methods based on the factorisation of the dominating measure following Fuchs (2013) closely.

For the sake of simplicity, consider the following $1-$dimensional SDE:

$$dX_t = b_\zeta(X_t)dt + \sigma_\zeta(X_t)dW_t, \tag{6.9}$$

with discrete observations without error so that $X_{t_i} = Y_{t_i}$, and $X_0 = x$. This setting can be considered as a missing data problem in the sense that we observe $X_0 = x$ and $X_T = x'$ but we cannot observe

the path between the two points so that we need to impute the missing path between the two. Write the measure induced by (6.9) with the conditions $X_0 = x$, $X' = x$ by $\mathbb{P}_{\zeta,x,x'}$. Also write $\mathbb{W}_\zeta$ to denote the law induced by $dX_t = \sigma_\zeta(X_t)dW_t$ on $[0,T]$, and $\mathbb{P}_\zeta$ be the law of (6.9) without conditions. Then Bayes' theorem gives rise to:

$$\frac{d\mathbb{P}_{\zeta,x,x'}}{d\mathbb{W}_{\zeta,x,x'}}(X) = \frac{\mathbb{P}_\zeta(X_T = x' \mid X)\mathbb{P}_\zeta(dX)/\mathbb{P}_\zeta(X_T \in dx')}{\mathbb{W}_\zeta(X_T = x' \mid X)\mathbb{W}_\zeta(dX)/\mathbb{W}_\zeta(X_T \in dx')}, \tag{6.10}$$

where we omit the reference to the initial position $X_0 = x$ and $\mathbb{W}_{\zeta,x,x'}$ denotes the law $\mathbb{W}_\zeta$ with $X_0 = x$, $X_T = x'$. From the Girsanov theorem (Theorem 6), we know the expression $\mathcal{G}(X) := \frac{\mathbb{P}_\zeta(dX)}{\mathbb{W}_\zeta(dX)}$, and the ratio $\frac{\mathbb{P}_\zeta(X_T=x'|X)}{\mathbb{W}_\zeta(X_T=x'|X)} = 1$ by definition. As a result, we have that:

$$\frac{d\mathbb{P}_{\zeta,x,x'}}{d\mathbb{W}_{\zeta x,x'}}(X) = \mathcal{G}(X)\frac{\mathbb{W}_\zeta(X_T \in dx' \mid X_0 = x)}{\mathbb{P}_\zeta(X_T \in dx' \mid X_0 = x)}. \tag{6.11}$$

Assume that $\mathbb{P}_\zeta(X_T \in dx' \mid X_0 = x)$ and $\mathbb{W}_\zeta(X_T \in dx' \mid X_0 = x)$ admit the transition densities $f_\theta(x' \mid x)$ and $q_\theta(x' \mid x)$ respectively. Then (6.11) becomes:

$$\frac{d\mathbb{P}_{\zeta,x,x'}}{d\mathbb{W}_{\zeta x,x'}}(X) = \mathcal{G}(X)\frac{q_\theta(x' \mid x)}{f_\theta(x' \mid x)}. \tag{6.12}$$

Although we have obtained a general expression of the problem, the expression in (6.12) is not really useful since it is not possible to obtain samples from it due to intractable densities in general. Also, as argued in Roberts and Stramer (2001), for any $\zeta \neq \zeta'$, $\mathbb{W}_\zeta$ and $\mathbb{W}_{\zeta'}$ are mutually singular measures so that they have disjoint support. Since $\mathbb{P}_\zeta$ and $\mathbb{W}_\zeta$ are equivalent, such singularity implies that, for any $\zeta \neq \zeta'$, $\mathbb{P}_\zeta$ and $\mathbb{P}_{\zeta'}$ are mutually singular measures as well. Therefore, we need to find the expression which is equivalent to (6.12) with parameter-free reference measure, that is a reference measure that should not depend on the parameters.

Write X_{mis} for the path of X excluding the observations $X_{obs} := \{X_{t_i}; 1 \leq i \leq N\} = \{Y_{t_i}; 1 \leq i \leq N\}$ thus $X = X_{mis} \cup X_{obs}$. It turns out that (6.12) can be used to simulate X_{mis} given X_{obs}. We then introduce a *Brownian bridge*. Let W_t be a standard Brownian motion. Then W_t conditioned on $W_s = x$ and $W_u = y$ is called a *Brownian* $(s,x,u,y)-$*bridge*. Then one can simulate a Brownian $(s,x,u,y)-$bridge with the following three steps.

i) Let $s = t_0 < t_1 < \cdots t_n = u$, and set $W_0 = 0$. Given σ, obtain samples according to $W_{i+1} \sim \mathcal{N}(W_i, \sigma^2\Delta_i)$ for $i = 0, \cdots n-1$ with $\Delta_i := t_{i+1} - t_i$.

ii) Obtain a Brownian $(s,0,u,0)-$bridge according to $\hat{W}_i = W_i - \frac{t_i - s}{u-s}W_i$ for $i = 0, \cdots n$.

iii) Then construct a Brownian $(s,x,u,y)-$bridge $\bar{W}$ as follows: $\bar{W}_i = \hat{W}_i + \frac{u-t_i}{u-s}x + \frac{t_i-s}{u-s}y$ for $i = 0, \cdots n$.

Assume that $\sigma_\zeta(X_t) = \sigma$, that is $\sigma_\zeta(X_t)$ is known and independent from X_t. This can be possible via the Lamperti transform (Proposition 9) for instance. Roberts and Stramer (2001) suggest the following reparameterisation, but with different motivation of ours, based on the above procedure to factorise $\mathbb{W}_\zeta$ as the product of a Brownian bridge. First we simulate a Brownian $(0,0,T,0)-$bridge with unit $\sigma = 1$,

and simulate x' from $\mathbb{W}_\zeta$ conditional on initial point x. Then transform the Brownian $(0,0,T,0)-$bridge to obtain a Brownian $(0,x,T,x')-$bridge with parameter σ as we described. Critically, the final transform can be considered as a map $W \mapsto \bar{W} := F_\zeta(W;x,x')$, where W denotes a standard Brownian motion, and $\bar{W}$ denotes a Brownian $(0,x,T,x')-$bridge. This implies that $\mathbb{W}_\zeta$ can be factorised as:

$$\mathbb{W}_\zeta\left(W_{[0,T)},x'\right) = \left(\mathbb{W}_\zeta^{(0,x,T,x')} \otimes \mathbb{W}_\zeta\right)\left(F_\zeta^{-1}\left(\bar{W};x,x'\right),x'\right), \tag{6.13}$$

where $\mathbb{W}_\zeta^{(0,x,T,x')}$ denotes the law induce by a Brownian $(0,x,T,x')-$bridge. Based on (6.13), Roberts and Stramer (2001) study the following practical decomposition of $\mathbb{P}_\zeta$. First consider the following transformations:

$$\begin{cases} \hat{X}_t := F_\zeta^{(1)}(X_t) = \frac{1}{\sigma}X_t, \\ \bar{X}_t := F_\zeta^{(2)}(\hat{X}_t,x,x') = \hat{X}_t - \frac{(T-t)\dot{x}+t\dot{x}'}{T}, \end{cases} \tag{6.14}$$

where $\dot{x} = x/\sigma$, $\dot{x}' = x'/\sigma$. From Ito's lemma, we have that:

$$d\hat{X}_t = \frac{b_\zeta(\hat{X}_t)}{\sigma}dt + dW_t, \tag{6.15}$$

with $\hat{X}_0 = \frac{x_0}{\sigma}$ so that $F_\zeta^{(1)}(\cdot)$ acts as the Lamperti transform. It can be easily seen that $\bar{X}_0 = 0$ and $\bar{X}_T = 0$ hold, thus $F_\zeta^{(2)}(\cdot)$ transforms $\hat{X}_t$ to start and finish at zero, that is now $\bar{X}_t$ is a Brownian $(0,0,T,0)-$bridge. Clearly $F_\zeta^{(1)}(\cdot)$ and $F_\zeta^{(2)}(\cdot)$ have the same role in (6.13), therefore we have that:

$$\bar{W} = F_\zeta(W;x,x') = F_\zeta^{(2)}\left(F_\zeta^{(1)}(W);F_\zeta^{(1)}(x),F_\zeta^{(1)}(x')\right),$$

and this gives rise to:

$$\frac{d\mathbb{P}_\zeta}{d\mathbb{W}_\zeta}(X_{mis} \mid x,x') = \frac{d\bar{\mathbb{P}}_\zeta}{d\mathbb{W}^{(0,0,T,0)}}(\bar{X}_{mis} \mid x,x'), \tag{6.16}$$

sot that the reference measure can be written independently of σ given x', where $\bar{\mathbb{P}}_\zeta$ denotes the law of the process $\bar{X}_t$ and $\mathbb{W}^{(0,0,T,0)}$ denotes the one of a Brownian $(0,0,T,0)-$bridge. Besides, from (6.16), we have that:

$$\begin{aligned} \frac{d\mathbb{P}_\zeta}{d\mathbb{W}_\zeta}(X_{mis} \mid x,x') &= \frac{d\hat{\mathbb{P}}_\zeta}{d\mathbb{W}^{(0,\bar{x},T,\bar{x}')}}(\hat{X}_{mis} \mid \bar{x},\bar{x}'), \\ &= \mathcal{G}_\zeta\left(\hat{X}_{[0,t]}\right), \end{aligned} \tag{6.17}$$

where $\hat{\mathbb{P}}_\zeta$ denotes the law of the process $\hat{X}_t$ in (6.15) so that the Girsanov term $\mathcal{G}_\zeta\left(\hat{X}_{[0,t]}\right)$ does not depend on σ. We note that this approach is, in general, only applicable to $1-$dimensional SDEs since it is well known that the transform $F_\zeta^{(1)}(\cdot)$ does generally not exist, see Ait-Sahalia (2008).

Kalogeropoulos et al. (2010) study a different approach based on a random time change technique (Øksendal, 2003, Chapter 8.2). Again consider (6.9) with general $\sigma_\zeta(X_t)$ where $t \in [0,1]$. Then

consider the following random time change function:

$$\beta(t) = \int_0^t c(s)ds, \tag{6.18}$$

for $t \in [0,1]$, where $c : [0,1] \to \mathbb{R}_+$ is called time change rate. In particular, we set $c(s) = \sigma^2$ so that $s := \beta_1(t) = \sigma^2 t$. Using this, define:

$$U_s = \begin{cases} X_{\beta_1^{-1}(t)}, \ 0 \leq \sigma^2 \leq s, \\ M_{\beta_1^{-1}(t)}, \ s > \sigma^2, \end{cases}$$

where $dM_t = \sigma dW_t$ so that $U_{\sigma^2} = X_1 = x'$. Applying Ito's lemma provides

$$dU_t = \begin{cases} \frac{b_\theta(U_s)}{\sigma^2}ds + dW_s, \ \ 0 \leq \sigma^2 \leq s, \\ dW_s, \ , \ s > \sigma^2. \end{cases}$$

Let $\mathbb{U}$ and $\mathbb{W}_\zeta^U$ denote the measure induced by U_t and $\sigma_\theta(U_t)dt + dW_t$. From the Girsanov's theorem (Theorem 6), we have:

$$\frac{d\mathbb{U}}{d\left(\mathbb{W}_\zeta^U \otimes \text{Leb}\right)}(U_{mis}, x, x') = \exp\left(\int_0^{\sigma^2} \frac{b_\theta(U_t)}{\sigma^2}dU_t - \frac{1}{2}\int_0^{\sigma^2} \frac{b_\theta(U_t)^2}{\sigma^4}dt\right) f_\zeta(x' \mid x),$$

where the reference measure $\mathbb{W}_\zeta^U \otimes \text{Leb}$ still depends on σ so that $\mathbb{W}_\zeta^U \otimes \text{Leb}$ is not still appropriate. This leads us to introduce a second time change transformation:

$$\begin{aligned} u = \beta_2(s) &= \frac{s}{\sigma^2(\sigma^2 - s)}, \\ \longleftrightarrow s = \beta_2^{-1}(u) &= \frac{\sigma^4 u}{1 + \sigma^2 u}. \end{aligned}$$

for $s \in [0, \sigma^2)$, and define a new process Z:

$$\begin{aligned} U_s &= (\sigma^2 - s)Z_{\beta_2(s)} + \left(1 - \frac{s}{\sigma^2}\right)x + \frac{s}{\sigma^2}x', \\ \longleftrightarrow Z_u &= \frac{1}{(\sigma^2 - s)}\left(U_s - \left(1 - \frac{s}{\sigma^2}\right)x + \frac{s}{\sigma^2}x'\right), \end{aligned}$$

for $u \in [0, \infty)$. Set $x = x' = 0$, then Z_u becomes:

$$Z_u = \frac{1 + u\sigma^2}{\sigma^2}U_{\beta_2^{-1}(u)}.$$

Applying Ito's lemma and the time change formula Øksendal (2003, Theorem 8.5.6) to Z_u gives rise to:

$$dZ_u = \left\{\frac{b_\zeta\left(\frac{\sigma^2 Z_u}{1+u\sigma^2}\right) + \sigma^2 Z_u}{1 + u\sigma^2}\right\}du + dW_u,$$

for $u \in [0, \infty)$. This operation essentially transforms to a diffusion that runs from 0 to ∞ preserving the unit volatility. Let $\mathbb{Z}$ denote the law of Z and $\mathbb{W}^Z$ denote the corresponding law of a unit diffusion process. Then we have that:

$$\frac{d\mathbb{Z}}{d\left(\mathbb{W}^Z \times \text{Leb}\right)}\left(Z_{[0,\infty)}, x, x'\right) \propto \mathcal{G}_\zeta\left(Z_{[0,\infty)}\right) \tag{6.19}$$

Critically, Kalogeropoulos et al. (2010, Corollary 3.1) show that the process Z is standard Brownian motion under $\mathbb{W}^Z$. It turns out that the reference measure $\mathbb{W}^Z \times \text{Leb}$ does not depend on σ so that one can do inference for Z based on the expression in (6.19) via MCMC for instance. We note that the approach described in Kalogeropoulos et al. (2010) cannot be applied to multidimensional SDEs in general although they have generalised for some multidimensional stochastic volatility models.

Remark 15. *Another drawback of the methods described in Roberts and Stramer (2001); Kalogeropoulos et al. (2010) is that they have shown equivalent expressions to* (6.12) *which are known up to the normalising constant because they focus on Bayesian inference via MCMC. Thus it is not straightforward to apply their methods to our problem since we need to an (equivalent) analytical expression of* $f_\theta(x' \mid x)$.

6.3 Forward-only smoothing

For convenience, here we again introduce Del Moral et al. (2010), see also subsubsection 4.5.1. The bootstrap filter Gordon et al. (1993) is immediately applicable in the continuous-time setting, as it only requires forward sampling of the underlying signal $X = \{X_t; t \geq 0\}$; this is trivially possible – under numerous approaches – and is typically associated with the introduction of some time-discretisation bias. However, the transition density is still required for the smoothing problem we have posed in the Introduction. In this section, we assume a standard discrete-time HMM, with transition density $f_\theta(x' \mid x)$, and potential function $g_\theta(y \mid x)$, for appropriate $x, x' \in \mathbb{R}^{d_x}$, $y \in \mathbb{R}^y$, and review an online algorithm in Del Moral et al. (2010) for this setting.

Implementation of the bootstrap filter (Algorithm 14) provides an immediate approximation of the smoothing distribution $p_\theta(dx_{0:n} \mid y_{0:n})$ by following the genealogy of the particles. This method is studied,e.g., in Cappe (2009); Dahlhaus and Neddermeyer (2010). Let $\{x_{0:n}^{(i)}, W_n^{(i)}\}_{i=1}^N$, $N \geq 1$, be a particle approximation of the smoothing distribution $p_\theta(dx_{0:n} \mid y_{0:n})$, in the sense that we have the estimate:

$$\begin{cases} \hat{p}_\theta(dx_{0:n} \mid y_{0:n}) & := \sum_{i=1}^N W_n^{(i)} \delta_{x_{0:n}^{(i)}}(dx_{0:n}), \\ \sum_{i=1}^N W_n^{(i)} & = 1, \end{cases} \tag{6.20}$$

where $\delta_{x_{0:n}^{(i)}}(dx_{0:n})$ is the Dirac measure with an atom at $x_{0:n}^{(i)}$. Then, replacing $p_\theta(dx_{0:n} \mid y_{0:n})$ with its estimate $\sum_{i=1}^N W_n^{(i)} \delta_{x_{0:n}^{(i)}}(dx_{0:n})$ provide consistent estimators of the quantity of interest $\mathcal{S}_{\theta,n}$ in (Algorithm 14).Though the method is online and the computational cost is $\mathcal{O}(N)$, it typically suffers from the well-documented path degeneracy problem – as illustrated via theoretical results or numerically (Del Moral et al., 2010; Kantas et al., 2015). That is, as n increases, the particles representing $p_\theta(dx_{0:n} \mid y_{0:n})$ obtained by the above method will eventually all share the same ancestral particle due to the resampling steps, thus the approximation collapses as $n \to \infty$.

An approach overcoming path-degeneracy is the Forward Filtering Backward Smoothing (FFBS) algorithm of Doucet et al. (2000). We briefly review the method here, following closely the notation and development in Del Moral et al. (2010) . In the forward direction, assume that a filtering algorithm (e.g. bootstrap) has provided a particle approximation of the filtering distribution $p_\theta(dx_{k-1} \mid y_{0:k-1})$, assuming a relevant k:

$$\hat{p}_\theta(dx_{k-1} \mid y_{0:k-1}) = \sum_{i=1}^{N} W_n^{(i)} \delta_{x_{k-1}^{(i)}}(dx_{0k-1}), \tag{6.21}$$

or weighted particles $\{x_{k-1}^{(i)}, W_k^{(i)}\}_{i=1}^N$. In the backward direction, assume that one is given the particle approximation of the marginal smoothing distribution $p_\theta(dx_k \mid y_{0:n})$:

$$\hat{p}_\theta(dx_k \mid y_{0:n}) = \sum_{i=1}^{N} W_{k|n}^{(i)} \delta_{x_k^{(i)}}(dx_k). \tag{6.22}$$

Then one has that (Kitagawa, 1987):

$$\begin{aligned} p_\theta(x_{k-1:k} \mid y_{0:n}) &= p_\theta(dx_k \mid y_{0:n}) \otimes p_\theta(dx_{k-1} \mid x_k, y_{0:k-1}), \\ &= p_\theta(dx_k \mid y_{0:n}) \otimes p_\theta(dx_{k-1} \mid y_{0:k-1}) \frac{f_\theta(x_k \mid x_{k-1})}{\int f_\theta(x_k \mid x_{k-1}) p_\theta(x_{k-1} \mid y_{0:k-1}) dx_{k-1}}. \end{aligned} \tag{6.23}$$

Using (6.21) , (6.22) and (6.23), we obtain the approximation:

$$\hat{p}_\theta(dx_{k-1:k} \mid y_{0:n}) = \sum_{j=1}^{N} W_{k|n}^{(j)} \sum_{i=1}^{N} \frac{f_\theta(x_k^{(j)} \mid x_{k-1}^{(i)}) W_{k-1}^{(i)}}{\sum_{l=1}^{N} f_\theta(x_k^{(j)} \mid x_{k-1}^{(l)}) W_{k-1}^{(l)}} \delta_{(x_{k-1}^{(i)}, x_k^{(j)})} dx_{k-1:k}. \tag{6.24}$$

Recalling the expectation of additive functionals in (6.6)-(6.7), the above calculations give rise to the following estimator of the target quantity $\mathcal{S}_{\theta,n}$ in (Algorithm 14):

$$\hat{\mathcal{S}}_{\theta,n} = \sum_{k=0}^{n} \int s_{\theta,k}(x_{k-1}, x_k) \hat{p}_\theta(dx_{k-1:k} \mid y_{0:n}).$$

To be able to apply the above method, the marginal smoothing approximation in (6.22) is obtained via a backward recursive approach. In particular, starting from $k = n$ (where the approximations provided by the standard forward particle filter), one proceeds as follows. Given k, the quantity for $k-1$ is directly obtained by integrating out x_k in (6.24) thus we have:

$$\hat{p}_\theta(dx_{k-1} \mid y_{0:n}) = \sum_{i=1}^{N} W_{k-1|n}^{(i)} \delta_{x_{k-1}^{(i)}}(dx_{k-1}),$$

for the normalised weights:

$$W_{k-1|n}^{(i)} \propto \sum_{j=1}^{N} W_{k|n}^{(j)} \frac{f_\theta(x_k^{(j)} \mid x_{k-1}^{(i)}) W_{k-1}^{(i)}}{\sum_{l=1}^{N} f_\theta(x_k^{(j)} \mid x_{k-1}^{(l)}) W_{k-1}^{(l)}}.$$

Notice that – in this version of FFBS – the same particles $\{x_k^{(i)}\}_{i=1}^N$ are used in both directions (the ones before resampling at the forward filter), but with different weights.

An important development made in Del Moral et al. (2010) is transforming the above offline algorithm into an online one. This is achieved by consideration of the sequence of instrumental functionals:

$$T_{\theta,0}(x_0) := s_{\theta,0}(x_0),\ T_{\theta,n}(x_n) := \int S_{\theta,n}(x_{0:n}) p_\theta(dx_{0:n-1} \mid y_{0:n-1}, x_n), \tag{6.25}$$

for $n \geq 1$. Notice that

$$\mathcal{S}_{\theta,n} = \int T_{\theta,n}(x_{0:n}) p_\theta(dx_n \mid y_{0:n}).$$

Then Del Moral et al. (2010, Proposition 2.1) show that:

$$\begin{aligned} T_{\theta,n}(x_n) &= \int \left[T_{\theta,n-1}(x_{0:n-1}) + s_{\theta,n}(x_{n-1}, x_n)\right] p_\theta(dx_{n-1} \mid y_{0:n-1}, x_n), \\ &= \frac{\int \left[T_{\theta,n-1}(x_{0:n-1}) + s_{\theta,n}(x_{n-1}, x_n)\right] f_\theta(x_n \mid x_{n-1}) p(dx_{n-1} \mid y_{0:n-1})}{\int f_\theta(x_n \mid x_{n-1}) p(dx_{n-1} \mid y_{0:n-1})}. \end{aligned} \tag{6.26}$$

This recursion provides an online – forward-only – advancement of FFBS for estimating the smoothing expectation of additive functionals. The complete method is summarised in Algorithm 17: the key ingredients that, during the recursion, values of the functional $T_{\theta,n}(x_n)$ are only required at the discrete positions $x_n^{(i)}$ determined by the forward particle filter.

In the SDE context, under Assumption 14, the transition density $f_\theta(x' \mid x)$ is considered intractable, thus Algorithm 17, apart from serving as a review of the method in Del Moral et al. (2010), does not appear to be practical in the continuous-time case.

6.4 Data augmentation on diffusion pathspace

To overcome the intractability of the transition density $f_\theta(x' \mid x)$ of the SDE, we will work with an algorithm that is defined in continuous-time and makes use of the complete SDE path-particles in its development. The new method has connections with earlier attempts in the literature Särkkä and Sottinen (2008) focus on the filtering problem for a class of models related to (6.1)-(6.1) and come up with an approach that requires the complete SDE path, for a limited class of SDEs with additive noise and no jumps. Fearnhead et al. (2008) also deal with the filtering problem, and – equipped with an unbiased estimator of the unknown transition density – recast the problem as one of filtering over an augmented space that incorporates the randomness for the unbiased estimate. Unfortunately, the method is accompanied by strict conditions on the drift and diffusion coefficient (the SDE – no jumps – can be transformed into one of unit diffusion coefficient and a drift that has a gradient form). Our contribution requires, in principle, solely the diffusion coefficient invertibility Assumption 15; arguably, the weakened condition we require is due to the fact that our approach appears as the relatively most natural extension (compared to alternative methods) of the 'standard' discrete-time algorithm of Del Moral et al. (2010).

The latter discrete-time method requires the density $f_\theta(x' \mid x) = f_\theta(dx' \mid x)/dx'$. In continuous-

time, we obtain an analytically available Radon-Nikodym derivative of $p_\theta(d\mathbf{x}' \mid x)$, for a properly defined variate $\mathbf{x}'$ that involves information about the continuous-time path for moving from x to x' within time Δ. We will give the complete algorithm. In this section, we prepare the ground via carefully determining $\mathbf{x}'$ given x, and calculating the relevant densities to be plugged in into the method.

6.4.1 SDEs with continuous paths

We work first with the process with continuous sample-paths, i.e. of dynamics:

$$dX_t = b_\zeta(X_t)dt + \sigma_\zeta(X_t)dW_t. \tag{6.27}$$

We adopt an approach motivated by techniques used for MCMC algorithms (Golightly and Wilkinson, 2008; Roberts and Stramer, 2001; Chib et al., 2004). Assume we are given starting point $x \in \mathbb{R}^{d_X}$, ending point point x', and the complete continuous-time path for the signal process in (6.27) on $[0, T]$, for some $T \geq 0$. That is, we now work with the path process:

$$[\{X_t; 0 \leq t \leq T\} \mid X_0 = x, X_T = x']. \tag{6.28}$$

Let $\mathbb{P}_{\zeta,x,x'}$ denote the law of path space-valued variable in (6.28). We consider the bridge process $\tilde{X} = \{\tilde{X}_t; 0 \leq t \leq T\}$ defined as:

$$d\tilde{X}_t = \left\{ b_\zeta(\tilde{X}_t) + \frac{x' - \tilde{X}_t}{T - t} \right\} dt + \sigma_\zeta(X_t)dW_t, \tag{6.29}$$

with $\tilde{X}_0 = x$ for $t \in [0, T]$. We denote the law of $\tilde{X}$ by $\mathbb{Q}_{\zeta,x,x'}$. Critically, a path of $\tilde{X}$ starts at point x and finishes at x', w.p.1. Under regularity conditions, Delyon and Hu (2006) prove that probability measures $\mathbb{P}_{\zeta,x,x'}$, $\mathbb{Q}_{\zeta,x,x'}$ are absolutely continuous w.r.t. each other.

We treat the auxiliary SDE (6.29) as a mapping from the driving noise to the solution, whence a sample path, X, of the process $\tilde{X} = \{\tilde{X}_t; 0 \leq t \leq T\}$, is produced by a mapping – determined by (6.29) – of a corresponding sample path, say Z, of the Wiener process. That is, we have defined a map, and – under Assumption 15 – its inverse:

$$Z \mapsto X := F_\zeta(Z; x, x'),\ Z = F_\zeta^{-1}(X; x, x'). \tag{6.30}$$

More analytically, F_ζ^{-1} is is defined via the transform:

$$dZ_t = \sigma_\zeta(X_t)^{-1} \left\{ dX_t - b_\zeta(X_t)dt - \frac{x' - X_t}{T - t}dt \right\}. \tag{6.31}$$

In this case, we define:

$$\mathbf{x}' := (x', Z),$$

and the probability measure of interest is:

$$p_\theta(d\mathbf{x}' \mid x) := f_\theta(dx' \mid x) \otimes p_\theta(dZ \mid x', x). \tag{6.32}$$

Then, we have the following results.

Lemma 12. *Let $\mathbb{Z}_{\zeta x,x'}$ be the law of Z when $X \sim \mathbb{P}_{\zeta,x,x'}$. Let $\mathbb{W}$ be the standard Wiener measure on $[0,T]$. Then we have that $\mathbb{Z}_{\zeta x,x'}$ is absolutely continuous w.r.t. $\mathbb{W}$.*

Proof. Theorem 6 of Delyon and Hu (2006) shows that $\mathbb{P}_{\zeta,x,x'}$ and $\mathbb{Q}_{\zeta,x,x'}$ are equivalent, and both X and Z are constructed via the map $F_\zeta(\cdot)$ in (6.30). This implies that if $X \sim \mathbb{Q}_{\zeta,x,x'}$ then $Z \sim \mathbb{W}$ thus the result follows. □

Lemma 13. *The process $F_\zeta\left(F_\zeta^{-1}(X;x,x')\right)$ hits the end points x' w.p.1 for $X \sim \mathbb{P}_{\zeta,x,x'}$.*

Proof. Lemma 12 ensures that $\mathbb{Z}_{\zeta x,x'}$ is absolutely continuous w.r.t. $\mathbb{W}$ so that the measure induced by $F_\zeta(Z;x,x')$ is also absolutely continuous w.r.t. the measure $\mathbb{Q}_{\zeta,x,x'}$ which is induced by $F_\zeta(W;x,x')$ where W is a Wiener process on $[0,T]$. The result follows from that the diffusion process hits the desired end point x' w.p.1 under $\mathbb{Q}_{\zeta,x,x'}$. □

Again, let $\mathbb{W}$ be the standard Wiener measure on $[0,T]$. Since we have:

$$\frac{p_\theta(dZ \mid x', x)}{\mathbb{W}(dZ)} = \frac{d\mathbb{P}_{\zeta,x,x'}}{d\mathbb{Q}_{\zeta,x,x'}}\left(F_\zeta(Z;x,x')\right),$$

remains to obtain the density $\frac{d\mathbb{P}_{\zeta,x,x'}}{d\mathbb{Q}_{\zeta,x,x'}}$. Such a Radom-Nikodym derivative has been object of interest in many works. Delyon and Hu (2006) provide detailed conditions and a proof, but (seemingly) omit an expression for the normalising constant which is important in our case, as it involves the parameter θ, in our applications later in the paper, we aim to infer about θ. Papaspiliopoulos and Roberts (2009); Papaspiliopoulos et al. (2013) provide an expression based on a conditioning argument for the projection of the probability measures on $[0,t)$, $t \leq T$ and passage to the limit $t \uparrow T$. The derivations in Delyon and Hu (2006) are extremely rigorous, so we will make use the expressions in that paper. Following carefully the proofs of some of their main results (Theorem 5, together with Lemmas 7, 8) one can indeed retrieve the constant in the deduced density. In particular, Delyon and Hu (2006) a impose the following conditions.

Assumption 16. *i) $v \mapsto \sigma_\zeta(v)$ is twice continuously differentiable and bounded, with bounded first and second derivatives, and it is invertible, with bounded inverse.*

ii) $v \mapsto b_\zeta(v)$ is locally Lipschitz, locally bounded.

iii) SDE (6.27) admits a strong solution.

Proposition 37. *Under Assumption 16 we have that:*

$$\frac{d\mathbb{P}_{\zeta,x,x'}}{d\mathbb{Q}_{\zeta,x,x'}}(F_\zeta(Z;x,x')) = \frac{\mid \Sigma_\zeta(x') \mid^{1/2}}{\mid \Sigma_\zeta(x) \mid^{1/2}} \times \frac{\mathcal{N}\left(x';x,T\Sigma_\zeta(x)\right)}{f_\theta(x' \mid x)} \times \mathcal{G}(X;x,x'),$$

where $\mathcal{G}(X;x,x')$ *is defined as:*

$$\log \mathcal{G}(X;x,x') := \int_0^T \left\langle b_\zeta(X_t), \Sigma_\zeta^{-1}(X_t) dX_t \right\rangle - \frac{1}{2}\int_0^T \left\langle b_\zeta(X_t), \Sigma_\zeta^{-1}(X_t) b_\zeta(X_t) dt \right\rangle$$
$$-\frac{1}{2}\int_0^T \frac{\left\langle (x'-X_t), d\Sigma_\zeta^{-1}(X_t)(x'-X_t)\right\rangle}{T-t} - \frac{1}{2}\frac{\sum_{i,j=1}^{d_x} d\left[\Sigma^{-1}_{\zeta,i,j}, (x'_i - X_{t,i})(x'_j - X_{t,j})\right]}{T-t}, \tag{6.33}$$

here $[\cdot,\cdot]$ *denotes the quadratic variation process for semi-martingales.*

We note that different transforms to (6.31) have been proposed in the literature (Dellaportas et al., 2006; Kalogeropoulos et al., 2010) to achieve the same effect of obtaining an $1-1$ mapping of the latent path that has a density w.r.t. a measure that does not depend on x, x' or ζ. However, such methods are typically applicable for scalar diffusions, see e.g. Ait-Sahalia (2008). Auxiliary variables involving a random, finite selection of points of the latent path, based on the (generalised) Poisson estimator of Fearnhead et al. (2008) are similarly restrictive. In contrast to other attempts, our methodology may be applied for a more general class of SDEs, as determined by Assumption 15, and further regularity conditions, as in Assumption 16.

As a consequence of Proposition 37, we have obtained that:

$$\frac{p_\theta(d\mathbf{x}' \mid x)}{\left(\mathrm{Leb}^{\otimes d_x} \otimes \mathbb{W}\right)(d\mathbf{x}')} = \mathcal{G}(X;x.x') \times \mathcal{N}\left(x';x,T\Sigma_\zeta(x)\right) \times \frac{\mid \Sigma_\zeta(x') \mid^{1/2}}{\mid \Sigma_\zeta(x) \mid^{1/2}} =: p_\theta(\mathbf{x}' \mid x). \tag{6.34}$$

Remark 16. *A critical point here is that the above density is analytically tractable, thus by working on pathspace we have overcome the unavailability of the transition density* $f_\theta(x' \mid x)$.

6.4.2 SDEs with jumps

We extend the above developments to the more general case of the d_x-dimensional jump diffusion model given in (6.1), which we re-write here for convenience:

$$dX_t = b_\zeta(X_{t-})dt + \sigma_\zeta(X_{t-})dW_t + dJ_t, \tag{6.35}$$

with $X_0 = x$, $t \in [0,T]$.

Recall that $J = \{J_t\}$ denotes a compound Poisson process with jump intensity $\lambda_\nu(\cdot)$ and jump-size density $h_\mu(\cdot)$; also $\eta = (\nu,\mu)$ and $\theta = (\zeta,\eta)$. Let $\mathbb{F}_{\theta,x}$ denote the law of the unconditional process (6.35) and $\mathbb{L}_\eta$ the law of the involved compound Poisson process. We write $J = ((\tau_1,b_1),\cdots(\tau_\kappa,b_\kappa))$ to denote the jump process, where $\{\tau_i\}$ are the times of events, $\{b_i\}$ are the jump-sizes and $\kappa > 0$ is the total number of events. In addition, we consider the reference measure $\mathbb{L}$, corresponding to unit rate Poisson process measure on $[0,T]$ multiplied with $\otimes_{i=1}^{\kappa+1}\mathrm{Leb}^{\otimes d_x}$.

Construct One We consider the random variate:

$$\mathbf{x}' = \left(J, \{x_{\tau_i-}\}_{i=1}^{\kappa+1}, \{Z(i)\}_{i=1}^{\kappa+1}\right),$$

under the conventions $x_{\tau_0-} := x$, $x_{\tau_{\kappa+1}} := x'$, where we have defined:

$$Z(i) := F_\zeta^{-1}\left(X(i); x_{\tau_{i-1}}, x_{\tau_i-}\right), \; X(i) := \left\{X_t; x_{\tau_{i-1}} \leq t \leq x_{\tau_i}\right\},$$

for $1 \leq i \leq \kappa + 1$. We have that:

$$p_\theta(d\mathbf{x}' \mid x) := \mathbb{L}_\eta(dJ) \otimes \left[\otimes_{i=1}^{\kappa+1} \left\{f_\theta(dx_{\tau_i-} \mid x_{\tau_{i-1}}) \otimes p_\theta(dZ(i) \mid x_{\tau_{i-1}}, x_{\tau_i-})\right\}\right].$$

Using the results about SDEs without jumps in subsubsection 6.4.1 upon defiining:

$$\mathbf{x}'(i) := (x_{\tau_{i^-}}, Z(i)),$$

we have that:

$$\frac{f_\theta(dx_{\tau_i-} \mid x_{\tau_{i-1}}) \otimes p_\theta(dZ(i) \mid x_{\tau_{i-1}}, x_{\tau_i-})}{\mathrm{Leb}^{\otimes d_x}(dx_{\tau_i-}) \otimes \mathbb{W}(dZ(i))} = p_\theta(\mathbf{x}' \mid x_{\tau_{i^-}}; \tau_i - \tau_{i-1}).$$

Thus, the density of $p_\theta(d\mathbf{x}' \mid x)$ w.r.t. the reference measure:

$$\mu(d\mathbf{x}') := \mathbb{L}(dJ) \otimes \left[\otimes_{i=1}^{\kappa+1} \left\{\mathrm{Leb}^{\otimes d_x}(dx_{\tau_i-}) \otimes \mathbb{W}(dZ(i))\right\}\right],$$

is equal to:

$$\begin{aligned}\frac{p_\theta(d\mathbf{x}' \mid x)}{\mu(d\mathbf{x}')} = \frac{e^{-\int_0^{-1} \lambda_\nu(t)dt}}{e^{-1}} \times \prod_{i=1}^{\kappa} \{\lambda_\nu(\tau_i) h_\mu(b_i)\} \\ \times \prod_{i=1}^{\kappa+1} p_\theta(\mathbf{x}' \mid x_{\tau_{i^-}}; \tau_i - \tau_{i-1}).\end{aligned} \tag{6.36}$$

Construct Two We adopt an idea used – for a very different problem – in Gonçalves and Roberts (2014). Given $x, x' \in \mathbb{R}^{d_x}$, we define an auxiliary process $\tilde{X}_t$ as follows:

$$d\tilde{X}_t = \left\{b_\zeta(\tilde{X}_t) + \frac{x' - J_T - \tilde{X}_t + J_t}{T - t}\right\} dt + \sigma_\zeta(\tilde{X}_t) dW_t + dJ_t, \tag{6.37}$$

with $X_0 = x$ so that $\tilde{X}_T = x'$ w.p.1. As before, we view (6.37) as a transform, projecting a path, Z of the Wiener process and the compound process, J, onto a path, X, of the jump process. That is, we consider the map:

$$(J, Z) \mapsto X =: F_\zeta(J, Z; x, x'), \; (J, Z) = F_\zeta^{-1}(X; x, x'). \tag{6.38}$$

Notice that for the inverse transform, the $J-$part is obtained immediately, whereas for the $Z-$ part one uses the expression – well-defined due to Assumption 15:

$$dZ_t = \sigma_\zeta(X_t)^{-1} \left\{dX_t - dJ_t - b_\zeta(X_t)dt - \frac{x' - J_T - X_t + J_t}{T - t} dt\right\}. \tag{6.39}$$

We denote by $\bar{\mathbb{P}}_{\zeta,x,x'}$ the law of the original process in (6.35) conditionally on hitting x' at time T. Also, we denote the distribution on pathspace induced by (6.39) as $\bar{\mathbb{Q}}_{\zeta,x,x'}$. Now consider the variate:

$$\mathbf{x}' = (x', J, Z),$$

so that:

$$p_\theta(d\mathbf{x}' \mid x) := f_\theta(dx' \mid x) \otimes p_\theta\left(d(J,Z) \mid x, x'\right).$$

Lemma 14. *Let $\mathbb{Z}_{\zeta,x,x'}$ be the law of $(J,Z) = F_\zeta^{-1}(X; x, x')$ for $X \sim \bar{\mathbb{P}}_{\theta,x,x'}$ and $\mathbb{L}_\eta \otimes \mathbb{W}$ be the law of the involved compound Poisson process and a Wiener process. Then $\mathbb{Z}_{\zeta,x,x'}$ is absolutely continuous w.r.t. $\mathbb{L}_\eta \otimes \mathbb{W}$.*

Proof. The assumptions on J_t and Delyon and Hu (2006, Theorem 1) ensure that $\bar{\mathbb{P}}_{\zeta,x,x'}$ and $\bar{\mathbb{Q}}_{\zeta,x,x'}$ are equivalent. X and (J,Z) are constructed via the map $F_\zeta(\cdot)$ in (6.38). Then the results follows from the same line in Lemma 12. □

Lemma 15. *The process $F_\zeta\left(F_\zeta^{-1}(X; x, x')\right)$ hits the end points x' w.p.1 for $X \sim \bar{\mathbb{P}}_{\zeta,x,x'}$.*

Proof. This can be proven by the same argument in Lemma 13. □

Therefore, due to the employed $1-1$ transforms, we have that:

$$\frac{p_\theta(d\mathbf{x}' \mid x)}{\left(\mathrm{Leb}^{\otimes d_x} \otimes \mathbb{L}_\eta \otimes \mathbb{W}\right)(d\mathbf{x}')} = f_\theta(dx' \mid x) \otimes \frac{d\bar{\mathbb{P}}_{\zeta,x,x'}}{d\bar{\mathbb{Q}}_{\zeta,x,x'}}\left(F_\zeta(J,Z;x,x')\right).$$

Thus, using the parameter-free reference measure:

$$\mu(d\mathbf{x}') := \mathrm{Leb}^{\otimes d_x} \otimes \mathbb{L} \otimes \mathbb{W},$$

one obtains that:

$$\begin{aligned}\frac{p_\theta(d\mathbf{x}' \mid x)}{\mu(d\mathbf{x}')} = f_\theta(x' \mid x) \times \frac{e^{-\int_0^T \lambda_\nu(t)dt}}{e^{-T}T^\kappa} \times \prod_{i=1}^{\kappa} h_\mu(b_i) \\ \times \frac{d\bar{\mathbb{P}}_{\zeta,x,x'}}{d\bar{\mathbb{Q}}_{\zeta,x,x'}}\left(F_\zeta(J,Z;x,x')\right).\end{aligned} \tag{6.40}$$

Remark 17. Delyon and Hu (2006) *obtained the Radon-Nikodym derivative after a great amount of rigorous analysis. A similar development for the case of conditioned jump diffusions does not follow from their work, and can only be subject of dedicated research at the scale of a separate paper. This is beyond the scope of our work. In practice, one can proceed as follows. For grid size $M \geq 1$, and $\delta = T/M$, let $\bar{\mathbb{P}}^M_{\theta,x,x'}\left(X_\delta, \ldots, X_{(M-1)\delta} \mid X_0 = x, X_{M\delta} = x'\right)$ denote the time-discretised Lebesgue density of the $M-1$-positions of the conditioned diffusion with law $\bar{\mathbb{P}}_{\theta x,x'}$. Once (6.40) is obtained, a*

time-discretisation approach will give:

$$\begin{aligned}&\bar{\mathbb{P}}^M_{\theta,x,x'}\left(X_\delta,\ldots,X_{(M-1)\delta} \mid X_0 = x, X_{M\delta} = x'\right)\\ =\ &\frac{\bar{\mathbb{P}}^M_{\theta,x,x'}\left(X_\delta,\ldots,X_{(M-1)\delta}, X_{M\delta} = x' \mid X_0 = x\right)}{\bar{\mathbb{P}}^M_{\theta,x,x'}\left(, X_{M\delta} = x' \mid X_0 = x\right)}.\end{aligned}$$

In this time-discretised setting, $f_\theta(x' \mid x)$ will be replaced by $\bar{\mathbb{P}}^M_{\theta,x,x'}\left(X_{M\delta} = x' \mid X_0 = x\right)$. Thus, the intractable transition density over the complete time period will cancel out, and one is left with an explicit expression to use on a PC. Compared to the method in SDEs with continuous paths, and the Construct One in the current section, we do not have explicit theoretical evidence of a density on the pathspace. Yet, all numerical experiments we tried showed that the deduced algorithm was stable under mesh-refinement. We thus adopt the approach (or, conjecture) that the density in (6.40) exists, under assumptions, and can be obtained pending future research.

6.5 Forward-only smoothing for SDEs

6.5.1 Pathspace algorithm

We are ready to develop a forward-only particle smoothing method, under the scenario in (6.6)-(6.7) on the pathspace setting. We will work with the pairs of random elements:

$$(x_{k-1}, \mathbf{x}_k),\ 1 \le k \le n, \tag{6.41}$$

with $\mathbf{x}_k$ as defined in (subsection 6.4), i.e. containing pathspace elements. with x_k given by an 1-1 transform of $\{X_s; s \in [t_{k-1}, t_k]\}$, such that we can obtain a density for $p_\theta(\mathbf{x}_k \mid x_{k-1})$ w.r.t. a reference measure that does not involve θ. Recall that $p_\theta(\mathbf{x}_k \mid x_{k-1})$ denotes the probability law for the augmented variable $\mathbf{x}_k$ given x_{k-1}. We also write the corresponding density as:

$$p_\theta(\mathbf{x}_k \mid x_{k-1}) := \frac{p_\theta(d\mathbf{x}_k \mid x_{k-1})}{\mu(d\mathbf{x}_k)}.$$

The quantity of interest is now:

$$\mathcal{S}_{\theta,n}(\mathbf{x}_{0:n}) = \int S_\theta(\mathbf{x}_{0:n}) p_\theta(d\mathbf{x}_{0:n} \mid y_{0:n}), \tag{6.42}$$

with $n \ge 1$ for the class of additive functionals $S(\cdot)$ of the structure:

$$S_\theta(\mathbf{x}_{0:n}) = \sum_{k=0}^{n} s_{\theta,k}(x_{k-1}, \mathbf{x}_k), \tag{6.43}$$

under the convention that $x_{-1} := \emptyset$. Notice that we now allow $s_k(\cdot,\cdot)$ to be a function of x_{k-1} and $\mathbf{x}_k$; thus, $s_k(\cdot,\cdot)$ can potentially correspond to integrals, or other pathspace functionals. We will work with a transition density on the enlarged space of $\mathbf{x}_k$.

Similarly to the discrete-time case in Section 2, we define the functional:

$$T_{\theta,n}(\mathbf{x}_n) := \int S_\theta(\mathbf{x}_{0:n}) p_\theta(d\mathbf{x}_{0:n-1} \mid y_{0:n-1}, \mathbf{x}_n).$$

Proposition 38. *We have that:*

$$\mathcal{S}_{\theta,n}(\mathbf{x}_{0:n}) = \int T_{\theta,n}(\mathbf{x}_n) p_\theta(d\mathbf{x}_n \mid y_{0:n}).$$

Proof. We have the integral:

$$\int T_{\theta,n}(\mathbf{x}_n) p_\theta(d\mathbf{x}_n \mid y_{0:n}) = \int S_\theta(\mathbf{x}_{0:n}) p_\theta(d\mathbf{x}_{0:n-1} \mid y_{0:n-1}, \mathbf{x}_n) p_\theta(d\mathbf{x}_n \mid y_{0:n}).$$

Also, simple calculations give rise to:

$$\begin{aligned} p_\theta(\mathbf{x}_{0:n-1} \mid y_{0:n-1}, \mathbf{x}_n) &= p_\theta(d\mathbf{x}_{0:n-1} \mid y_{0:n}, \mathbf{x}_n), \\ &= \frac{p_\theta(\mathbf{x}_{0:n} \mid y_{0:n})}{p_\theta(\mathbf{x}_n \mid y_{0:n})}. \end{aligned}$$

Using this expression integral completes the proof. □

Critically, we obtain the following recursion. (We provide a proof for completeness.)

Proposition 39. *For any $n \geq 1$, we have that:*

$$\begin{aligned} T_{\theta,n}(\mathbf{x}_n) &= \int \left[T_{\theta,n-1}(\mathbf{x}_{n-1}) + s_{\theta,n}(x_{n-1}, \mathbf{x}_n)\right] p_\theta(d\mathbf{x}_{0:n-1} \mid y_{0:n-1}, \mathbf{x}_n), \\ &= \frac{\int \left[T_{\theta,n-1}(\mathbf{x}_{n-1}) + s_{\theta,n}(x_{n-1}, \mathbf{x}_n\right] p_\theta(\mathbf{x}_n \mid x_{n-1}) p_\theta(d\mathbf{x}_{n-1} \mid y_{0:n-1})}{\int p_\theta(\mathbf{x}_n \mid x_{n-1}) p_\theta(d\mathbf{x}_{n-1} \mid y_{0:n-1})}. \end{aligned}$$

Proof. Simply note that:

$$\begin{aligned} & p_\theta(d\mathbf{x}_{0:n-2} \mid y_{0:n-2}, \mathbf{x}_{n-1}) p_\theta(d\mathbf{x}_{n-1} \mid y_{0:n-1}, \mathbf{x}_n) \\ &= p_\theta(d\mathbf{x}_{0:n-2} \mid \mathbf{x}_{n-1}, y_{0:n-1}, \mathbf{x}_n) p_\theta(d\mathbf{x}_{n-1} \mid y_{0:n-1}, \mathbf{x}_n) \\ &= p_\theta(d\mathbf{x}_{0:n-1} \mid y_{0:n-1}, \mathbf{x}_n). \end{aligned}$$

Replacing the probability measure on the left side of the above equality with its equal on the right side, and using the latter in the integral above completes the proof for the first equation in the statement of the proposition. The second equation follows from trivial use of Bayes rule. □

Proposition 39 gives rise to a Monte-Carlo methodology for a forward-only, online approximation of the smoothing expectation of interest. This is given in Algorithm 23

Algorithm 23 Online Forward-Only Particle Smoothing on Pathspace

i) Initialise particles $\left\{x_0^{(i)}, W_0^{(i)}\right\}_{i=1}^N$, with $x_0^{(i)} \overset{i.i.d.}{\sim} \mu_\theta(dx_0)$, $W_0^{(i)} = g_\theta(y_0^{(i)} \mid x_0^{(i)})$, and functionals $\hat{T}_{\theta,0}(x_0^{(i)}) = s_{\theta,0}(\mathbf{x}_0^{(i)})$, for $1 \leq i \leq N$.

ii) Assume that at time $n-1$, one has a particle approximation $\left\{\mathbf{x}_{n-1}^{(i)}, W_{n-1}^{(i)}\right\}_{i=1}^N$ of the filtering law $p_\theta(d\mathbf{x}_{n-1} \mid y_{0:n-1})$ and estimators $\hat{T}_{\theta,n-1}(\mathbf{x}_{n-1}^{(i)})$ of $T_{\theta,n-1}(\mathbf{x}_{n-1})$, for $1 \leq i \leq N$.

iii) As time n, sample $\mathbf{x}_n^{(i)}$, for $1 \leq i \leq N$, from:

$$\mathbf{x}_n^{(i)} \sim \hat{p}_\theta(d\mathbf{x}_n \mid y_{0:n-1}) = \sum_{j=1}^N W_{n-1}^{(j)} p_\theta(d\mathbf{x}_n \mid x_{n-1}^{(j)}),$$

and assign particle weights $W_n^{(i)} \propto g_\theta(y_n \mid y_{n-1}, \mathcal{F}_n^{(i)})$, $1 \leq i \leq N$.

iv) Then, set, for $1 \leq i \leq N$:

$$\hat{T}_{\theta,n}(\mathbf{x}_n^{(i)}) = \frac{\sum_{j=1}^N W_{n-1}^{(j)} p_\theta(\mathbf{x}_n^{(i)} \mid x_{n-1}^{(j)})}{\sum_{l=1}^N W_{n-1}^{(l)} p_\theta(\mathbf{x}_n^{(i)} \mid x_{n-1}^{(l)})} \left[\hat{T}_{\theta,n-1}(\mathbf{x}_{n-1}^{(j)}) + s_{\theta,n}(x_{n-1}^{(j)}, \mathbf{x}_n^{(i)})\right].$$

v) Obtain an estimate of $\mathcal{S}_{\theta,n}$ as:

$$\hat{\mathcal{S}}_{\theta,n} = \sum_{i=1}^N W_n^{(i)} \hat{T}_{\theta,n}(\mathbf{x}_n^{(i)}).$$

Remark 18. *Although we have used the multinomial resampling, this should be understood as notational convenience. Indeed, other resampling methods such as the systematic resampling (Carpenter et al., 1999) and the stratified resampling (Kitagawa, 1996) can be directly used at the resampling step in Algorithm 23. We refer to Douc and Cappé (2005) for theoretical analysis of resampling methods in the context of SMC. Also notice that, our methodology can be applied to the auxiliary particle filter (Pitt and Shephard, 1999; Johansen and Doucet, 2008). Therefore, our developments are general enough to cover a broad range of SMC methods.*

6.5.2 Pathspace versus finite-dimensional construct

One can attempt to define an algorithm without reference to the underlying pathspace. That is, in the case of no jumps (for simplicity) an alternative approach can involve working with a regular grid on the period $[0,T]$, say $\{s_j = j\delta\}_{j=0}^M$, with $\delta = T/M$ for chosen size $M \geq 1$. Then, defining $\mathbf{x}' = (x_\delta, x_{2\delta}, \ldots, x_{M\delta})$, and using, e.g., an Euler-Maruyama time-discretisation scheme to obtain the joint density of such an $\mathbf{x}'$ given $x = x_0$, a Radon-Nikodym derivative, $p_\theta^M(\mathbf{x}'|x)$ on $\mathbb{R}^{M \times d_x}$, can be obtained with respect to the Lebesgue reference measure $\text{Leb}^{\otimes(d_x \times M)}$, as a product of M conditionally Gaussian densities. As shown e.g.˜in the motivating example in the Introduction, such an approach would lead to estimates with variability that increases rapidly with M, for fixed Monte-Carlo particles N. A central argument in this work is that one should develop the algorithm in a manner that respects

the probabilistic properties of the SDE pathspace, before applying (necessarily) a time-discretisation for implementation on a PC. This procedure is not followed for purposes of mathematical rigour, but it has practical effects on algorithmic performance.

6.5.3 Consistency

For completeness, we provide the following stability result of Algorithm 23. Consider the following assumptions.

Assumption 17. Let $\mathbf{X}$ and X denote the state spaces of $\mathbf{x}$ and x respectively.

i) For any relevant y', y, $\mathcal{F}$ and x, $g_\theta(y' \mid y, \mathcal{F}) = g_\theta(y \mid x)$ is a positive function such that $\sup_{x \in \mathsf{X}} \mid g_\theta(y \mid x) \mid < \infty$.

ii) $\sup_{\mathbf{x}' \in \mathbf{X}, x \in \mathsf{X}} |p_\theta(\mathbf{x}' \mid x)| < \infty$.

Proposition 40. *i) Under Assumption 17, for any $n \geq 0$, there exist constants $b_n, c_n > 0$ such that for any $\epsilon > 0$:*

$$\mathbb{P}\left(\left|\mathcal{S}_{\theta,n} - \hat{\mathcal{S}}_{\theta,n}\right| > \epsilon\right) \leq b_n e^{-c_n N \epsilon^2}.$$

ii) For any $n \geq 0$, $\hat{\mathcal{S}}_{\theta,n} \to \mathcal{S}_{\theta,n}$ w.p.1 as the number of particles $N \to \infty$.

Proof. The first part follows by the same arguments as Olsson and Westerborn (2017, Corollary 2). Assume that the first statement holds at time $n-1$. Define $\alpha_n := N^{-1} \sum_{i=1}^N g(y_n \mid x_n^{(i)})(T_n^N(\mathbf{x}_n^{(i)}) - \mathcal{S}_n(\theta))$ and $\beta_n := N^{-1} \sum_{i=1}^N g(y_n \mid x_n^{(i)})$ so that $\frac{\alpha_n}{\beta_n} = \mathcal{S}_{\theta,n} - \hat{\mathcal{S}}_{\theta,n}$. From the assumption, $\mid \beta_n \mid \leq \mid g_\theta(y_n \mid x) \mid_\infty$ holds so that:

$$\mathbb{P}\left(\mid \beta_n - \mathbb{E}\left[\beta_n \mid \mathcal{F}_{n-1}^N\right] \mid > \epsilon\right) \leq 2e^{-c_n N \epsilon^2}, \tag{6.44}$$

holds by Azuma (1967), where $\mathcal{F}_{n-1}^N$ denotes $\sigma-$algebra generated by Algorithm 23 at time $n-1$. By the construction, we obtain $\mathbb{E}\left[\beta_n \mid \mathcal{F}_{n-1}^N\right] = \sum_{i=1}^N W_{n-1}^{(i)} \int g_\theta(y_n \mid x_n) p_\theta(d\mathbf{x}_n \mid x_{n-1}^{(i)})$. From the induction assumption, we thus have that:

$$\mathbb{P}\left(\left|\mathbb{E}\left[\beta_n \mid \mathcal{F}_{n-1}^N\right] - \int \left[\int g_\theta(y_n \mid x_n) p_\theta(d\mathbf{x}_n \mid x_{n-1})\right] p_\theta(d\mathbf{x}_{n-1} \mid y_{0:n-1})\right| > \epsilon\right) \leq b_n e^{-c_n N \epsilon^2}. \tag{6.45}$$

Also by the assumption, we have $\left|g_\theta(y_n \mid x_n^{(i)}) \left(\hat{T}_{\theta,n}(\mathbf{x}_n^{(i)}) - \mathcal{S}_{\theta,n}\right)\right| \leq 2 \mid g_\theta(y_n \mid x) \mid_\infty \mid S_\theta(\mathbf{x}_{0:n}) \mid_\infty$. Thus:

$$\mathbb{P}\left(\mid \alpha_n - \mathbb{E}\left[\alpha_n \mid \mathcal{F}_{n-1}^N\right] \mid > \epsilon\right) \leq 2e^{-c_n N \epsilon^2},$$

follows from Azuma (1967). Then, again by the construction, we can show that $\mathbb{E}\left[\alpha_n \mid \mathcal{F}_{n-1}^N\right] = \sum_{i=1}^N W_{n-1}^{(i)} \int p_\theta(d\mathbf{x}_n \mid x_{n-1}^{(i)}) g_\theta(y_n \mid x_n) \left[\hat{T}_{\theta,n-1}(\mathbf{x}_{n-1}^{(i)}) + s_{\theta,n}(x_{n-1}^{(i)}, \mathbf{x}_n) - \mathcal{S}_{\theta,n}\right]$ and:

$$\int \left[\int p_\theta(d\mathbf{x}_n \mid x_{n-1}) g_\theta(y_n \mid x_n) \left[T_{\theta,n-1}(\mathbf{x}_{n-1}) + s_{\theta,n}(x_{n-1}, \mathbf{x}_n) - \mathcal{S}_{\theta,n}\right]\right] p_\theta(d\mathbf{x}_{n-1} \mid y_{0:n-1}) = 0.$$

Thus, the induction assumption gives rise to:

$$\mathbb{P}\left(\left|\mathbb{E}\left[\alpha_n \mid \mathcal{F}_{n-1}^N\right]\right| > \epsilon\right) \leq b_n e^{-c_n N \epsilon^2},$$

so that:

$$\mathbb{P}\left(\mid \alpha_n \mid > \epsilon\right) \leq b_n e^{-c_n N \epsilon^2}. \tag{6.46}$$

At time $n = 0$, we immediately have:

$$\mathbb{P}\left(\left|\mathcal{S}_{\theta,0} - \hat{\mathcal{S}}_{\theta,0}\right| > \epsilon\right) \leq b_0 e^{-c_0 N \epsilon^2}, \tag{6.47}$$

since $\hat{\mathcal{S}}_{\theta,0}$ is simply an importance sampling estimator and thus the standard Hoeffding's inequality can be applied. From (6.44)-(6.47), the first claim follows from Douc et al. (2011a, Lemma 4)).

Given $n \geq 0$, define $A_N(1/j) := \{\mid \mathcal{S}_{\theta,n} - \hat{\mathcal{S}}_{\theta,n}^N \mid > \frac{1}{j}\}$, here we have used N to emphasise the dependency of a particle estimate $\hat{\mathcal{S}}_{\theta,n}^N$ on N. Then we have:

$$\begin{aligned}\mathbb{P}\left(\lim_{N\to\infty} \hat{\mathcal{S}}_{\theta,n}^N = \mathcal{S}_{\theta,n}\right) &= 1 - \mathbb{P}\left(\cup_{j=1}^{\infty} \limsup_{N\to\infty} A_N(1/j)\right) \\ &\geq 1 - \sum_{j=1}^{\infty} \mathbb{P}\left(\limsup_{N\to\infty} A_N(1/j)\right).\end{aligned}$$

From this we have that $\mathbb{P}\left(A_N(1/j)\right) \leq b_n e^{-c_n N \left(\frac{1}{j}\right)^2}$ so that we obtain $\sum_{N=1}^{\infty} \mathbb{P}\left(A_N(1/j)\right) < \infty$. Therefore, $\mathbb{P}\left(\limsup_{N\to\infty} A_N(1/j)\right) = 0$ follows from the Borel–Cantelli lemma, and this gives rise to $\mathbb{P}\left(\lim_{N\to\infty} \hat{\mathcal{S}}_{\theta,n}^N = \mathcal{S}_{\theta,n}\right) = 1$. □

6.6 Online parameter/state estimation for SDEs

In this section, we derive an online gradient-ascent for partially observed SDEs. Following closely Poyiadjis et al. (2011); Del Moral et al. (2010), we focus on a computational method to estimate the MLE, that is, a gradient-ascent.

Poyiadjis et al. (2011) use the score function estimation methodology to propose an online gradient-ascent algorithm for obtaining an MLE-type parameter estimate, following ideas in Le Gland and Mevel (1997). In more detail, the method is based on the Robbins-Monro-type of recursion:

$$\begin{aligned}\theta_{n+1} &= \theta_n + \gamma_{n+1} \nabla \log p_{\theta_{0:n}}(y_n \mid y_{0:n-1}) \\ &= \theta_n + \gamma_{n+1} \left\{\nabla \log p_{\theta_{0:n}}(y_{0:n}) - \nabla \log p_{\theta_{0:n-1}}(y_{0:n-1})\right\},\end{aligned} \tag{6.48}$$

where $\{\gamma_n\}_{n\geq 1}$ is a positive decreasing sequence with:

$$\sum_{n=1}^{\infty} \gamma_n = \infty, \ \sum_{n=1}^{\infty} \gamma_n^2 < \infty.$$

The meaning of quantity $\nabla \log p_{\theta_{0:n}}(y_{0:n})$ is that – given a recursive method (in n) for the estimation of $\theta \mapsto \nabla \log p_{\theta_{0:n}}(y_{0:n})$ as we describe below and based on the methodology of Algorithm 23– one uses θ_{n-1} when incorporating y_{n-1}, then θ_n for y_n, and similarly for $k > n$. See LeGland and Mevel (1997); Tadic and Doucet (2018) for analytical studies of the convergence properties of the deduced algorithm, where under strong conditions the recursion is shown to converge to the 'true' parameter value, say $\theta_\star$, as $n \to \infty$.

Observe that, from Fisher's identity (see Poyiadjis et al. (2011); Cappé et al. (2005)) we have that:

$$\nabla \log p_\theta(y_{0:n}) = \int \nabla \log p_\theta(\mathbf{x}_{0:n}, y_{0:n}) p_\theta(d\mathbf{x}_{0:n} \mid y_{0:n}). \tag{6.49}$$

Thus, in the context of Algorithm 23, estimation of the score function corresponds to the choice:

$$\begin{cases} s_{\theta,k}(x_k, \mathbf{x}_{k-1}) & = \nabla \log p_\theta(\mathbf{x}_k, y_k \mid x_{k-1}), \\ S_\theta(\mathbf{x}_{0:n}) & = \nabla \log p_\theta(\mathbf{x}_{0:n}, y_{0:n}) = \sum_{k=0}^n \nabla \log p_\theta(\mathbf{x}_k, y_k \mid x_{k-1}). \end{cases}$$

Combination of the Robins-Morno recursion (6.48) with the one in Algorithm 23, delivers Algorithm 3, which we have presented here in some detail for clarity.

Algorithm 24 Online gradient-ascent for SDEs via forward-only smoothing

i) Assume that at time $n \geq 0$, one has a particle approximation $\left\{\mathbf{x}_n^{(i)}, W_n^{(i)}\right\}_{i=1}^N$ of the filtering law $p_{\hat{\theta}_{0:n}}(d\mathbf{x}_n \mid y_{0:n})$ and estimators $\hat{T}_{\hat{\theta}_{0:n},n}(\mathbf{x}_n^{(i)})$ of $T_{\hat{\theta}_{0:n},n}(\mathbf{x}_{n-1})$, for $1 \leq i \leq N$, and current parameter estimate $\hat{\theta}_n$

ii) Apply the iteration:

$$\hat{\theta}_{n+1} = \hat{\theta}_n \gamma_{n+1} \left\{ \nabla \log \widehat{p_{\hat{\theta}_{0:n}}}(y_{0:n}) - \nabla \log \widehat{p_{\hat{\theta}_{0:n-1}}}(y_{0:n-1}) \right\}.$$

iii) As time $n+1$, sample $\mathbf{x}_{n+1}^{(i)}$, for $1 \leq i \leq N$, from:

$$\mathbf{x}_{n+1}^{(i)} \sim \hat{p}_{\hat{\theta}_{n+1}}(d\mathbf{x}_{n+1} \mid y_{0:n}) = \sum_{j=1}^N W_n^{(j)} p_{\hat{\theta}_{n+1}}(d\mathbf{x}_{n+1} \mid x_n^{(j)}),$$

and assign particle weights $W_{n+1}^{(i)} \propto g_{\hat{\theta}_{n+1}}(y_{n+1} \mid y_n, \mathcal{F}_{n+1}^{(i)})$, $1 \leq i \leq N$.

iv) Then, set, for $1 \leq i \leq N$:

$$\hat{T}_{\hat{\theta}_{0:n+1},n+1}(\mathbf{x}_{n+1}^{(i)}) = \frac{\sum_{j=1}^N W_n^{(j)} p_{\hat{\theta}}(\mathbf{x}_{n+1}^{(i)} \mid x_n^{(j)})}{\sum_{l=1}^N W_n^{(l)} p_{\hat{\theta}}(\mathbf{x}_{n+1}^{(i)} \mid x_n^{(l)})} \left[\hat{T}_{\hat{\theta}_{0:n},n}(\mathbf{x}_n^{(j)}) + s_{\theta,n}(x_n^{(j)}, \mathbf{x}_{n+1}^{(i)}) \right],$$

where, on the right-hand-side we use the parameter $\theta = \hat{\theta}_{n+1}$.

v) Obtain an estimate as:

$$\hat{\mathcal{S}}_{\theta_{n+1},n+1} = \sum_{i=1}^N W_{n+1}^{(i)} \hat{T}_{\hat{\theta}_{0:n+1},n+1}(\mathbf{x}_{n+1}^{(i)}).$$

Remark 19. *When the joint density of* $(\mathbf{x}_{0:n}, y_{0:n})$ *is in the exponential family, an online EM algorithm can also be developed; see Del Moral et al. (2010) for the discrete-time case.*

6.7 Numerical Applications

When running the algorithms detailed below on a PC, we discretised the pathspace using the Euler-Maruyama scheme with $M = 10$ time points per unit of time; the cost of the algorithms scales linearly with M. To select the schedule of the tuning parameter $\{\gamma_k\}$ in (6.48), we adopt the well-known adaptive method developed in Kingma and Ba (2014) termed Adam to stabilise the unnecessary numerical instability due to the choice of $\{\gamma_k\}$. Let $c_n := -\left(\nabla \log p_{\hat{\theta}_{0:n}}(y_{0:n}) - \nabla \log p_{\hat{\theta}_{0:n-1}}(y_{0:n-1})\right)$.

Then, Adam involves the following iterative steps:

$$\begin{cases} m_n = m_{n-1}\beta_1 + (1-\beta_1)c_n \\ v_n = v_{n-1}\beta_2 + (1-\beta_2)c_n^2 \\ \hat{m}_n = m_n/(1-\beta_1^n), \\ \hat{v}_n = v_n/(1-\beta_2^n), \\ \hat{\theta}_{n+1} = \hat{\theta}_n - \alpha\hat{m}_n/\left(\sqrt{\hat{v}_n} + \epsilon\right), \end{cases} \tag{6.50}$$

where $(\beta_1, \beta_2, \alpha, \epsilon)$ are the tuning parameters. Theoretical and empirical convergence properties of Adam have been widely studied, see Goodfellow et al. (2016); Kingma and Ba (2014); Reddi et al. (2019) for instance. Following Kingma and Ba (2014), we set $(\beta_1, \beta_2, \alpha, \epsilon) = (0.9, 0.999, 0.001, 10^{-8})$ throughout the rest of the paper.

6.7.1 Ornstein-Uhlenbeck SDE

We consider the following model:

$$dX_t = -\theta_1 X_t + \theta_2 dW_t, \tag{6.51}$$

where we set $(\theta_1, \theta_2) = (0.4, 0.5)$, with observations $y_i = x_i + \epsilon_i$, $\epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, 0.1^2)$, $\Delta = 1$. Notice that one can calculate the score function of (6.51) analytically. To see the numerical stability of the algorithm we have developed, we simulated the data sets for $n = 2500$, 5000, 7500 and 10,000. Then we applied Algorithm 23 to approximate the score function with the number of particles $N = 50$, 100, and 150, each of these approximation experiments was replicated 50 times.

The results are plotted in Figure 13. The figure gives boxplots of the estimated score function values of θ_1, with the black dashed lines indicating the true values score function of. As one can see, the algorithm estimates the score function at the true parameter value quite accurately for large enough N. Based on the results, one can reasonably suggest that the asymptotic variance of the estimators is bounded uniformly over n, in agreement with Proposition 39 and Del Moral et al. (2015, Theorem 3.1).

We add an extra parameter, and work with the model:

$$dX_t = \theta_1(\theta_2 - X_t)dt + \theta_3 dW_t + dJ_t^\theta, \tag{6.52}$$

where again data consist of noisy observations, $y_i = X_i + \epsilon_i$ for $i = 1, 2, \cdots n$ where $\epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, 0.1^2)$ and $J_t^\theta = \sum_{i=1}^{N_t} \xi_i$ is a compound Poisson process that N_t is a Poisson process with intensity $\theta_4 t$ and $\xi_i \overset{i.i.d.}{\sim} \mathcal{U}(-\theta_5, \theta_5)$. We set the true parameter $(\theta_1^\star, \theta_2^\star, \theta_3^\star, \theta_4^\star, \theta_5^\star) = (0.3, 0.0, 0.2, 0.5, 0.5)$ for (6.52) with jumps and $(\theta_1^\star, \theta_2^\star, \theta_3^\star, \theta_4^\star, \theta_5^\star) = (0.2, 0.0, 0.2, 0.0, 0.0)$ (6.52) without jumps, and simulate $n = 20,000$ for both models. We then applied Algorithm 24 to these models given $(\theta_4^\star, \theta_5^\star)$ with $N = 100$ particles. The results are plotted in Figure 14.

To compare the efficiency of the construction 1 and construction 2, we applied Algorithm 23 based on each construction to approximate the score function with $N = 50, 100, 150, 200$ particles for the

data simulated according to (6.52) of size $n = 10$. Each experiment was replicated $R = 50$times with the parameter $(\theta_1^\star, \theta_2^\star, \theta_3^\star, \theta_4^\star, \theta_5^\star) = (0.3, 0.0, 0.2, 0.5, 0.5)$. The results are plotted in Figure 15. We also note that the computational cost of the construction 2 is much cheaper than that of the construction 1.

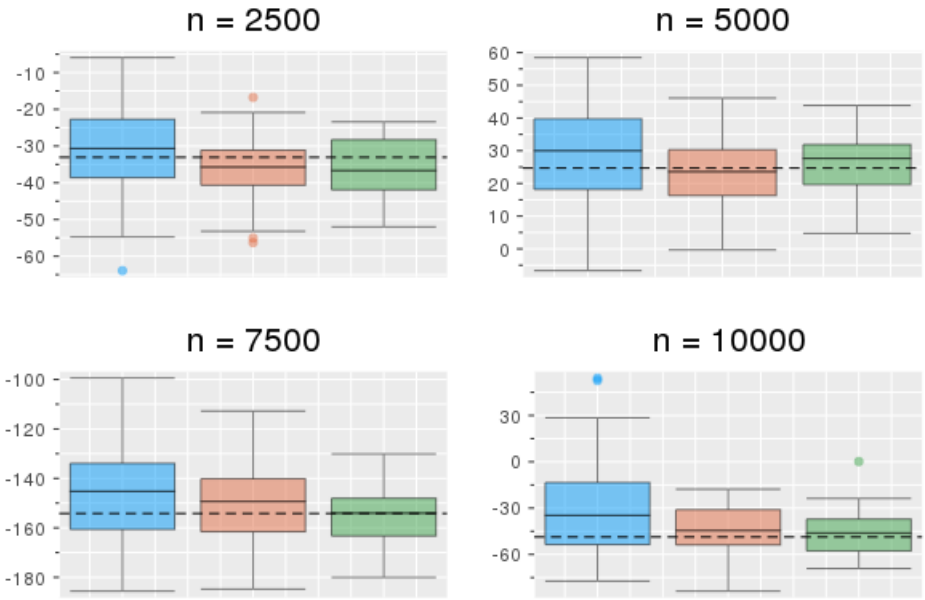


Figure 13: The boxplots of the estimated score function of θ_1 of the model in (6.51). We set $(\theta_1, \theta_2) = (0.4, 0.5)$ with observations $y_i = x_i + \epsilon_i$, $\epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, 0.1^2)$. The blue, orange and green box plots stand for the cases $N = 50$, 100 and 150 respectively. The black dash lines are the true values $(-33.1, 24.7, -154.2, -48.7)$ for $n = 2500$, 5000, 7500 and $10,000$ respectively.

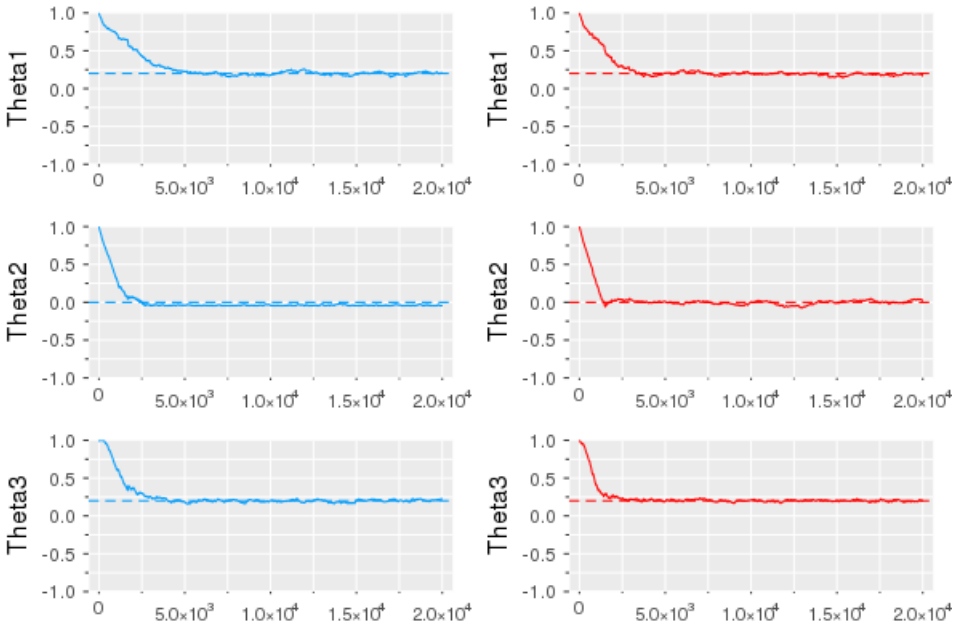


Figure 14: Trajectories from the online estimation of θ obtained from application of Algorithm 24 with $N = 100$ particles and initial value $(1.0, 1.0, 1.0)$ for $(\theta_1, \theta_2, \theta_3)$ respectively. Left panel shows the results for (6.52) with jumps, and the right one shows (6.52) without jumps. The horizontal dashed lines in the plots show the true parameter values $(\theta_1^\star, \theta_2^\star, \theta_3^\star) = (0.2, 0.0, 0.2)$. We set $(\theta_4^\star, \theta_5^\star) = (0.5, 0.5)$ for the jump model.

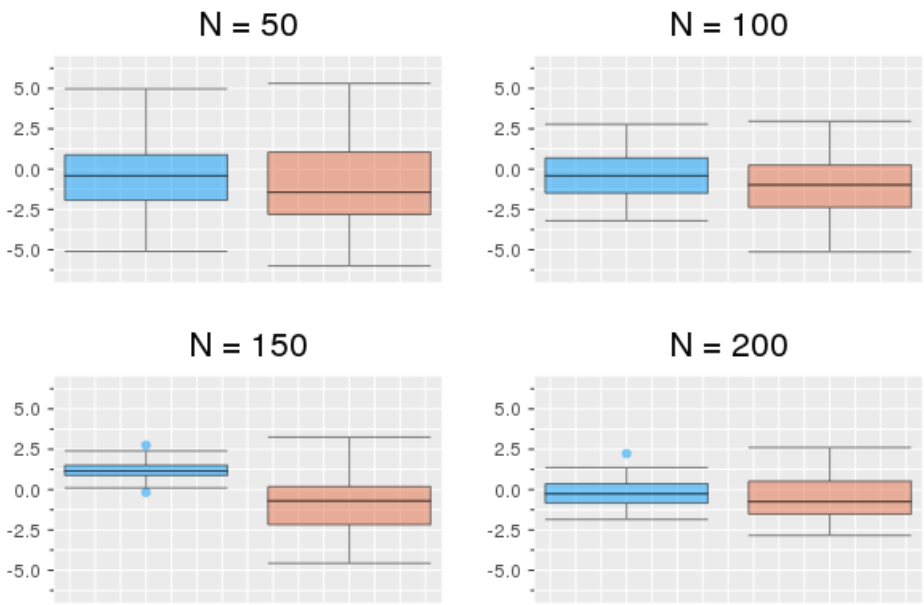


Figure 15: Boxplots of estimated score functions of θ_1 over $R = 50$ experiment repetitions, for model (6.52) with true parameter $(\theta_1^\star, \theta_2^\star, \theta_3^\star, \theta_4^\star, \theta_5^\star) = (0.3, 0.0, 0.2, 0.5, 0.5)$ with $n = 10$. Each orange boxplot corresponds to the construction 1 and blue one corresponds to the construction 2.

6.7.2 Periodic Drift SDE

Consider the following non-liner model:

$$dX_t = \sin\left(X_t - \theta_1\right) dt + \theta_2 dW_t, \tag{6.53}$$

where $0 \leq \zeta_1 \leq 2\pi$ and the data consist of noisy observations, $y_i = X_{t_i} + \epsilon_i$ where $\epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, 0.1^2)$. We set $(\theta_1^\star, \theta_2^\star) = (\pi/4, 0.9)$ as the true parameter. $(M, N, n, \tau) = (10, 100, 10000, 0.1)$ with the initial values $(\theta_1^0, \theta_2^0) = (0.1, 2)$. To simulate (6.53), we applied the Euler–Maruyama with the mesh $\delta = 1/M$. The results are plotted in Figure 16.

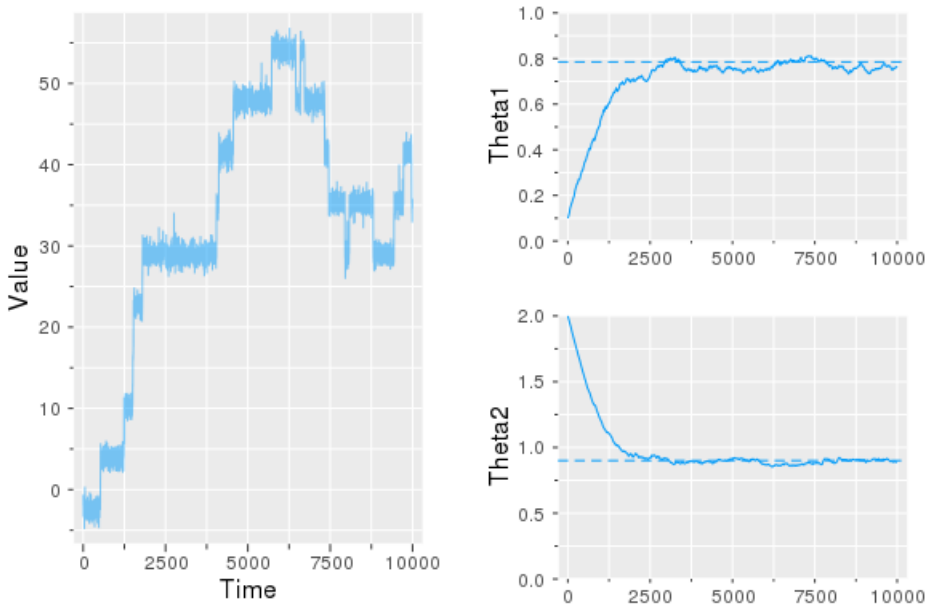


Figure 16: (*Left*) Data set simulated according to the sine diffusion observed with error with parameter values $(\theta_1, \theta_2) = (\pi/4, 0.9)$ in (6.53). The blue solid line indicates the values of state X_i. The observations were obtained with errors which were distributed according to $\mathcal{N}(0, 0.1^2)$. (*Right*) Online estimation of θ_1 (top) and θ_2 (bottom) for the data set. We set $(0.1, 2)$ as the initial values for (θ_1, θ_2) respectively with 100 particles in Algorithm 24. The horizontal dash lines indicate the true parameter values in each case.

6.7.3 Heston model

Following Heston (1993), suppose that the dynamics of the underlying asset's price S_t is given $dS_t = \theta_4 S_t dt + \sqrt{X_t} S_t dW_t^U$ where X_t is CIR process. Define log-prices $U_t = \log(S_t)$, and then Itô's lemma gives rise to:

$$\begin{cases} dU_t = \left(\theta_4 - \frac{X_t}{2}\right) dt + \sqrt{X_t} \left\{ \left(1 - \theta_5^2\right)^{1/2} dW_t + \theta_5 dB_t \right\}, \\ dX_t = \theta_1(\theta_2 - X_t)dt + \theta_3 \sqrt{X_t} dB_t, \end{cases} \tag{6.54}$$

where W_t and B_t are Brownian motions, $2\theta_1\theta_2 > \theta_3^2$ and $-1 \leq \theta_5 \leq 1$. We assume that U_t can be observed discretely so that $y_i = U_{t_i}$ for $i = 1, 2, \cdots n$. Under this setting, the likelihood function of observations y_i given y_{i-1}, $X_{t\in[t_{i-1},t_i]}$, θ is:

$$y_i \mid y_{i-1},\, X_{t\in[t_{i-1},t_i]},\, \theta \sim \mathcal{N}(y_i; \mu_i, \Sigma_i),$$

where:

$$\begin{cases} \mu_i = y_{i-1} + \int_{t_{i-1}}^{t_i} \left\{\theta_4 - \frac{1}{2}X_t\right\} dt + \theta_5 \int_{t_{i-1}}^{t_i} \sqrt{X_t} dB_t, \\ \Sigma_i = \left(1 - \theta_5^2\right) \int_{t_{i-1}}^{t_i} X_t dt. \end{cases} \tag{6.55}$$

We set $(\theta_1^\star, \theta_2^\star, \theta_3^\star, \theta_4^\star, \theta_5^\star) = (0.1, 1.0, 0.2, 0.45, 0.0)$ as the true parameter. We estimated the parameters via Algorithm 24 with the initial values $(\theta_1^0, \theta_2^0, \theta_3^0, \theta_4^0, \theta_5^0) = (0.005, 0.1, 0.4, 0.3)$ and fixed θ_5 as we conformed that estimating this parameter was difficult. We also set $(M, N, n, \tau) = (10, 100, 10000, 0.1)$. The results are given in Figure 17.

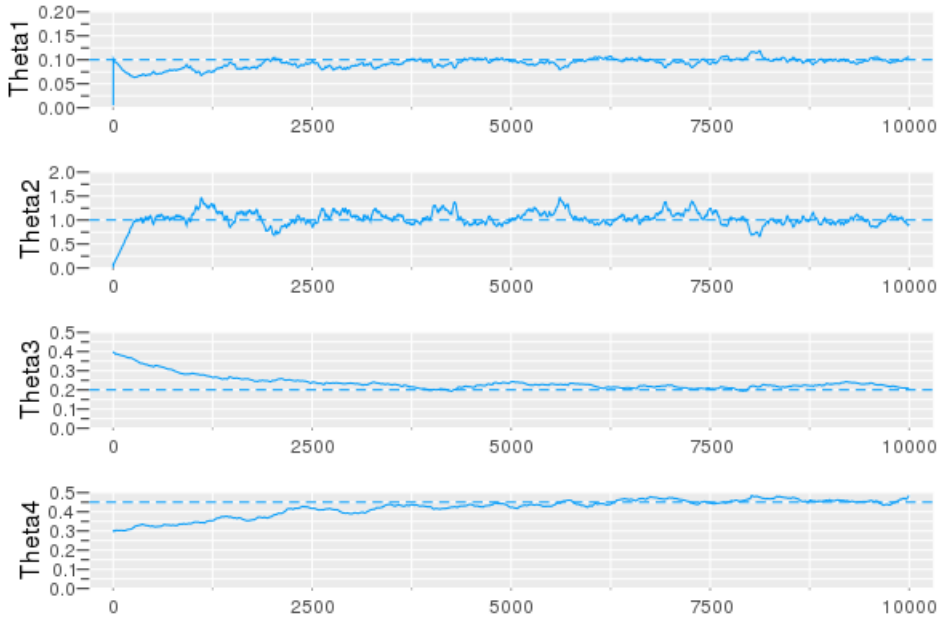


Figure 17: Online estimation of θ_1 (top), θ_2 (second top) , θ_3(second bottom) and θ_4 (bottom) for the data set simulated according to (6.54). We set $(0.005, 0.1, 0.4, 0.3)$ as the initial values for $(\theta_1, \theta_2, \theta_3, \theta_4)$ respectively with 100 particles in Algorithm 24. The horizontal dash lines indicate the true parameter values in each case.

6.7.4 Applications to real data with sequential model selection

We apply the method we have developed to real data with focus on model selection, motivated by Eraker et al. (2003); Johannes et al. (2009). Suppose that now one has a set of candidate models

$\{\mathcal{M}_k\}_{k\in\mathcal{K}}$. Then Bayesian Information Criterion (BIC) Schwarz (1978) of the model $\mathcal{M}_k$ is given by:

$$BIC(\mathcal{M}_k) := -2\ell_{\hat{\theta}_n^k}(y_{0:n-1}) + dim(\theta_k)\log n, \tag{6.56}$$

here $\hat{\theta}_n^k$ stands for the MLE of the model $\mathcal{M}_k$, and one selects the model minimising (6.56) as the best model among the models being considered. BIC can be obtained as the $1-$st order Laplace approximation of the log marginal likelihood (or evidence), and thus the difference:

$$BIC(\mathcal{M}_{l,k}) := -2\left(\ell_{\hat{\theta}_n^l}(y_{0:n-1}) + \ell_{\hat{\theta}_n^k}(y_{0:n-1})\right) + \log n\left(dim(\theta_l) - dim(\theta_k)\right), \tag{6.57}$$

could be considered an approximation of the negative log Bayes factor (Kass and Raftery, 1995). Since the (log) Bayes factor is strongly consistent in many cases (Chib and Kuffner, 2016), BIC is also (Nishii, 1988; Sin and White, 1996). Critically, this is also the case in the context of discrete time HMMs (Yonekura et al., 2018) so that it may be also true for our setting since the model being considered is essentially continuous time HMM. To be precise, assume that the models being considered are nested in the sense that a sequence of nested parametric models $\mathcal{M}_1 \subset \cdots \mathcal{M}_k \subset \cdots \mathcal{M}_p$ is specified via a sequence of corresponding parameter spaces $\Theta^1 \subseteq \mathbb{R}^{d_1}$, and $\Theta^{k+1} = \Theta^k \times \Delta\Theta^k$, $\Delta\Theta^k \subseteq \mathbb{R}^{d_k+1}$, $k \geq 1$ with $dim(\theta_k) < dim(\theta_l)$. In this case, if the model $\mathcal{M}_l$ is the true one, then $BIC(\mathcal{M}_{l,k}) \to -\infty$ as $n \to \infty$ w.p.1. Or, if the model $\mathcal{M}_k$ is the true one, then $BIC(\mathcal{M}_{l,k}) \to \infty$ as $n \to \infty$ w.p.1. See also Eguchi and Masuda (2018) for a rigorous analysis of BIC for diffusion-type models. Hereafter we always assume that $dim(\theta_k) < dim(\theta_l)$ for $BIC(\mathcal{M}_{l,k})$ where $k < l$.

It is worth noting that although Johannes et al. (2009) use the sequential likelihood ratio for such model comparison, this quantity might be overshooting. That is, the likelihood ratio might tend to choose a large model. Therefore, due to the penalty term, using the ratio of BICs might be more sensible than using the likelihood ratio for the sake of identifying a model. Also, they use fixed calibrated parameters so that they do not make statistical inference. Besides, Eraker et al. (2003) use the Bayes factor for model selection. Since they use MCMC to estimate parameters, their approach is not sequential. In contrast, the method we have studied allows us to estimate sequentially (also online) parameters. Therefore, our approach might be understood as a generalization of Johannes et al. (2009); Eraker et al. (2003).

Remark 20. *In our setting, we want to use the numerical studies inspired by the problems in finance to illustrate empirical performance and availability of the algorithm for real data so that we will use the outcome of the online recursion as a proxy for the MLE. Then BIC will be approximated by running our method for the chosen MLE value to obtain an approximation of the log-likelihood of the data at this parameter value. Therefore, obtained BIC has to be understood as a proxy of* (6.56).

Short-term interest rates We consider the following models for short-term interest rates:

$$dX_t = b_\theta^{(i)}(X_t) + \theta_4\sqrt{X_t}dW_t, \tag{6.58}$$

where:

$$\begin{cases} \mathcal{M}_1: & b_\theta^{(1)}(X_t) = \theta_0 + \theta_1 X_t, \\ \mathcal{M}_2: & b_\theta^{(2)}(X_t) = \theta_0 + \theta_1 X_t + \theta_2^2 X_t^2, \\ \mathcal{M}_3: & b_\theta^{(3)}(X_t) = \theta_0 + \theta_1 X_t + \theta_2^2 X_t^2 + \frac{\theta_3}{X_t}. \end{cases} \tag{6.59}$$

This class of model has been used routinely to study non-linearity of drift in the short-term interest rates, we refer to Jones (2003); Durham (2003); Ait-Sahalia (1996); **?** and references therein for a more in-depth treatment. Motivated by Dellaportas et al. (2006); Stanton (1997), we applied our method to the 3-month treasury bill rates which can be obtained from FRED, Federal Reserve Bank of St.Louis; `https://fred.stlouisfed.org/series/TB3MS`. We studied daily data from January 2 1970 to December 29, 2000 which consist of 7739 observations. The data set is plotted in Figure 18.

To obtain reasonable results, we first estimated the parameters with some arbitrary initial values, and calculated the mean of the estimates after a burn-in period of 1000 time steps for each parameter. Then we again estimated each parameter using the obtained mean as the initial values for estimating the parameters. Also, we used the mean of the data 6.62 for the initial value X_0 of the algorithm. We did the same procedure for each model.

The estimated result are given in Figure 19, Figure 20 and Figure 21 for $\mathcal{M}_1$, $\mathcal{M}_2$ and $\mathcal{M}_3$ respectively.

Figure 22 shows BIC difference in (6.57) for each pair. Recall that, for $dim(\theta_k) < dim(\theta_l)$ where $k < l$, if the model $\mathcal{M}_l$ is the true model, then $BIC(\mathcal{M}_{l,k}) \to -\infty$ as $n \to \infty$ and vice versa. First of all, the result implies that the model $\mathcal{M}_2$ did not the fit the data among the models during the period. Next, the model $\mathcal{M}_3$ might be appropriate one especially after 1994 but the model $\mathcal{M}_1$ might also adequately fit the date before 1994. These results imply that, on average, models with a non-liner drift function might be better to use to model the daily data of 3-month treasury bill rates, but the evidence is not that strong in agreement with the empirical studies in Chapman and Pearson (2000); Durham (2003); Dellaportas et al. (2006).

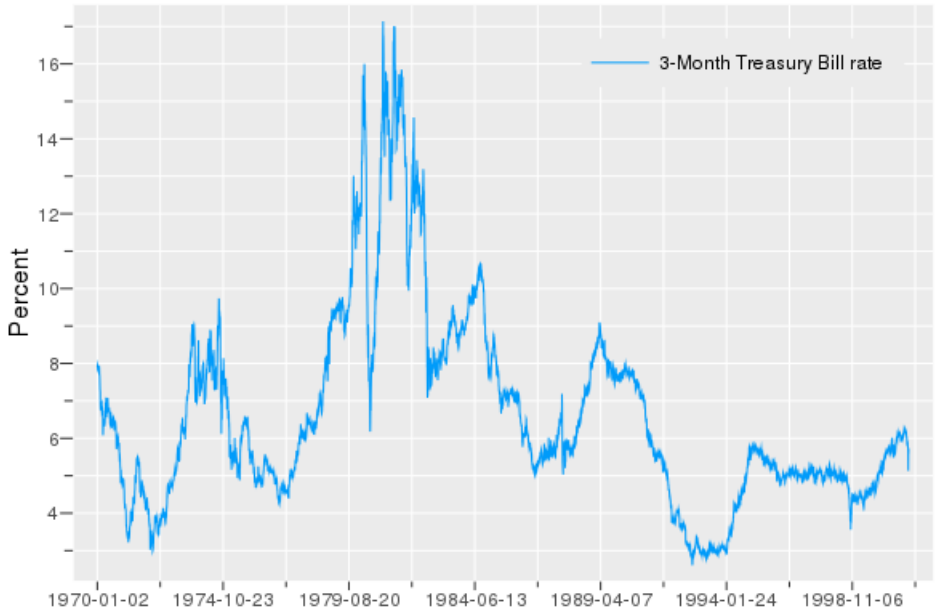
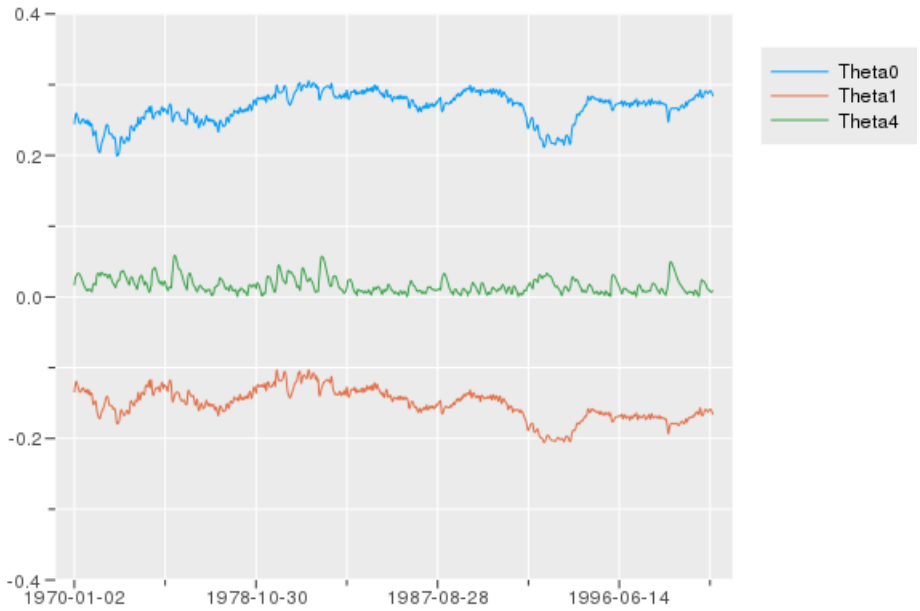


Figure 18: The daily data of the 3-month Treasury Bill rates from January 2, 1970 to December 29, 2000.



approcash

Figure 19: Online estimation o the model $\mathcal{M}_1$ for the data set in Figure 18. We set $(0.243, -0.136, 0.0153)$ as the initial values for $(\theta_0, \theta_1, \theta_4)$ respectively with 100 particles in Algorithm 24.

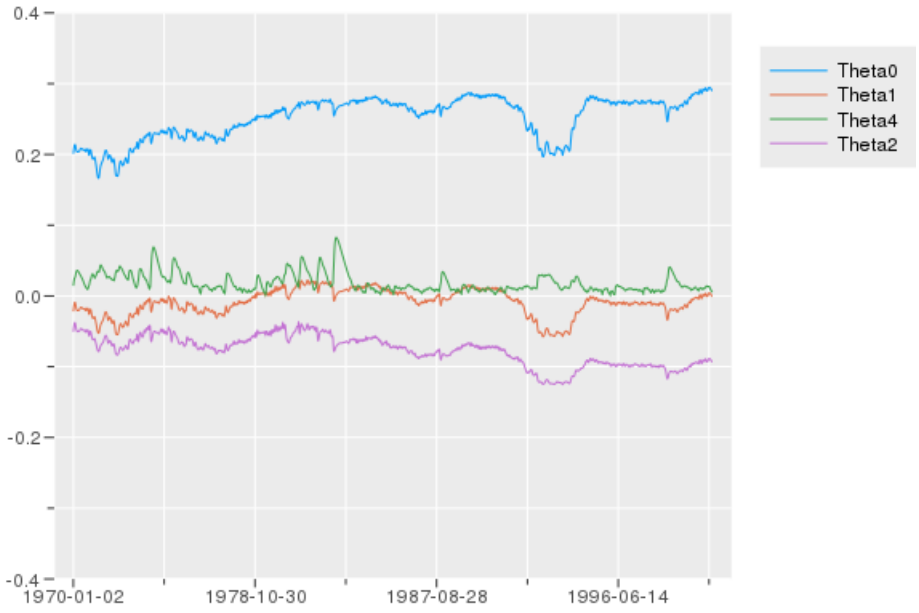


Figure 20: Online estimation of the model $\mathcal{M}_2$ for the data set in Figure 18. We set $(0.259, -0.0064, -0.079, 0.017)$ as the initial values for $(\theta_0, \theta_1, \theta_2, \theta_4)$ respectively with 100 particles in Algorithm 24.

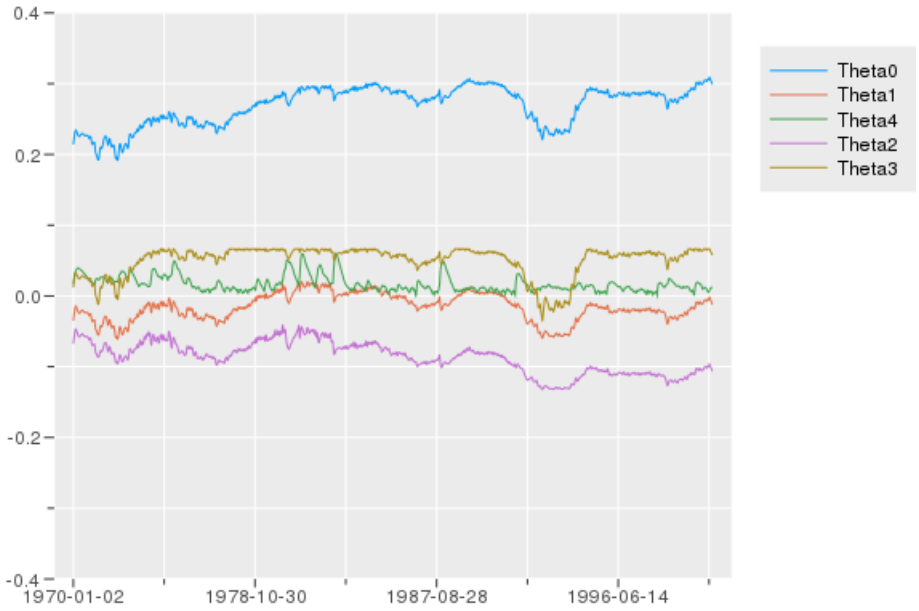


Figure 21: Online estimation of the model $\mathcal{M}_3$ for the data set in Figure 18. We set $(0.21, -0.036, -0.067, 0.011, 0.016)$ as the initial values for $(\theta_0, \theta_1, \theta_2, \theta_3, \theta_4)$ respectively with 100 particles in Algorithm 24.

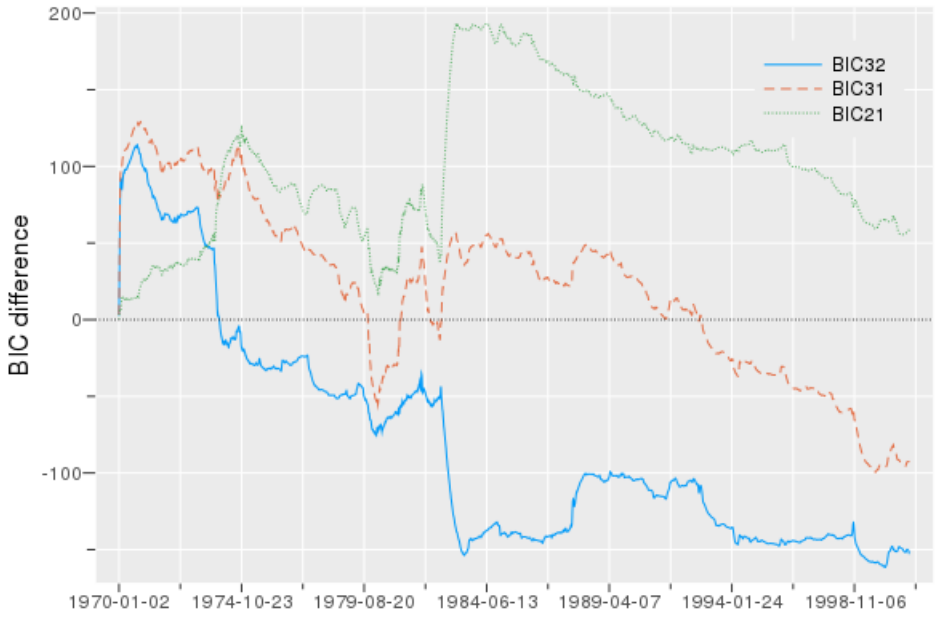


Figure 22: Online estimation of BIC difference defined in (6.57) for each model. The green solid line stands for $BIC(\mathcal{M}_{32})$, the orange dash line stands for $BIC(\mathcal{M}_{31})$, and the light blue dot stands for $BIC(\mathcal{M}_{21})$.

6.8 Conclusion and remarks

We have introduced an online particle smoothing methodology for discretely observed (jump) diffusions with intractable transition densities. Our approach overcomes such intractability by formulating the problem on pathspace, thus delivering an algorithm that -- besides regulatory conditions -- requires only the weak invertibility Assumption (Assumption 15). Thus, we have covered a rich family of SDE models, when related literature imposes strong restrictions. Combining our online smoothing algorithm with a Robbins-Monro-type approach of Recursive Maximum-Likelihood, we set up an online stochastic gradient-ascent for the likelihood function of the SDEs under consideration. The algorithm provides a wealth of interesting output, that can provide a lot of useful insights in statistical applications. The numerical examples show a lot of promise for the performance of the methodology. Our framework opens up a number of routes for insights and future research, including the ones described below.

i) In the case of SDEs of jumps, we have focused on jump dynamics driven by compound Poisson processes. There is great scope for generalisation here, and one can extend the algorithm to different cases of jump processes, also characterised by more complex dependencies between the jumps and the paths of the solution of the SDE, $X = \{X_t\}$. Extensions to time-inhomogeneous cases are immediate; we have chosen the time-homogeneous models only for purposes of presentation simplicity. The method can also be easily adopted to models with continuous-time data, once such information is separated in blocks of time intervals of $T = \mathcal{O}(1)$ length, notice that the incremental score function splits into a signal component (where all the pathspace construct will be applied) and a component involving the data given the signal, that, in principle, can be

of any form without effect on the derivation of the algorithm.

ii) Since the seminal work of Delyon and Hu (2006), more 'tuned' auxiliary bridge processes have appeared in the literature, see e.g. the works of Schauer et al. (2017); van der Meulen and Schauer (2017). Indicatively, the work in Schauer et al. (2017) considers bridges of the form (in one of the many options they consider) $d\tilde{X}_t = \left\{ b(\tilde{X}_t) + \Sigma^{-1}(x')\Sigma(\tilde{X})\frac{x'-\tilde{X}_t}{T-t} \right\} dt + \sigma(\tilde{X}_t)dW_t$. Auxiliary bridge processes that are closer in dynamics to the diffusion bridges of the given signal are expected to reduce the variability of Monte-Carlo algorithm, thus progress along the above direction can be immediately incorporated in our methodology and improve its performance. For instance, as noted in Schauer et al. (2017), use of the auxiliary bridge processes will give a Radon-Nikodym derivative where stochastic integrals cancel out. Such a setting is known to considerably reduce the variability of Monte-Carlo methods, see e.g.˜the numerical examples in Durham and Gallant (2002) and the discussion in Papaspiliopoulos and Roberts (2009, Section 4).

iii) The exact specification of the recursion used for the online estimation of unknown parameters is in itself an problem of intensive research in the field of stochastic optimisation. One would ideally aim for the recursion procedure to provide parameter estimates which are as close to the unknown parameter as the data (considered thus far) permit. In our case, we have used a fairly 'vanilla' recursion, maybe with the exception of the Adam variation. E.g., recent works in the Machine Learning community have pointed at the use of 'velocity' components in the recursion to speed up convergence, see, e.g., Sutskever et al. (2013); Yuan et al. (2016).

iv) We have mentioned through the main text several modifications that can improve algorithmic performance: dynamic resampling, stratified resampling, non-blind proposals in the filtering steps, choice of auxiliary processes. Parallelisation and use of HPC are obvious additions in this list.

v) Finally, we stress that the algorithm involves a filtering step, and a step that approximates the values of the instrumental function. These two procedures should be thought of separately. A reason of including two approaches in the case of jump diffusions (Constructs 1 and 2) is indeed to highlight this point. The two Constructs are identical in terms of the filtering part. Construct 1 incorporates in $\mathbf{x}'$ the location of the path at all times of jumps; thus, when the algorithm 'mixes' all pairs of $\{x_{k-1}^{(j)}\}$, $\{\mathbf{x}_k^{(i)}\}$, at the update of the instrumental function (see Step (iv) of Algorithm 23), many of such pairs can be incompatible. Such an effect is event stronger in the case of the standard algorithm applied in the motivating example in the Introduction, and partially explains the inefficiency of that algorithm. In contrast, in Construct 2, $\mathbf{x}'$ contains less information about the underlying paths, thus improving the compatibility of pairs selected from particle populations $\{x_{k-1}^{(j)}\}$, $\{\mathbf{x}_k^{(i)}\}$, thus, not surprisingly, Construct 2 seems more effective than Construct 1.

7 Adaptive Bayesian Model Selection for Diffusion Models

7.1 Introduction

This section studies MCMC methods for Bayesian model selection on high dimensional spaces. In particular, we focus on diffusion models potentially driven by fractional Brownian motion. Our approach is well defined on the diffusion path space so that the mixing rate of the algorithm might not depend on the mesh of discretization. Also, our new approach can avoid successfully a problem arising from the nature of trans-dimensional MCMC methods, as a consequence of the exact approximation of the posterior model probability. Critically, we propose the usage of MCMC methods on high dimensional spaces within sequential Monte Carlo samplers to learn adaptively some tuning parameters of the algorithm.

Let $(E, \mathcal{E})$ be a mesurable space, and $\mathcal{P}(E)$ denote the set of all probability measures on this space. Suppose that we are interested in sampling from $\eta(dx) \in \mathcal{P}(E)$ such that:

$$\eta(dx) := \frac{\mathcal{L}(x)\pi_0(dx)}{\mathcal{Z}}, \tag{7.1}$$

where $\mathcal{L} : E \to \mathbb{R}$ is a likelihood function, $\pi_0(dx) \in \mathcal{P}(E)$ is a prior distribution, and $\mathcal{Z} := \int_E \mathcal{L}(x)\pi_0(dx)$ is a marginal likelihood function so that $\eta(dx)$ corresponds to a posterior distribution. The main objective of this study is to develope a Markov chain Monte Carlo method to estimate $\mathcal{Z}$ on a high dimensional space.

$\mathcal{Z}$ is also called the *evidence* (Jeffreys, 1998) of the model in the context of the Bayesian model selection due to the followings. Assume that one has two candidate models, say $\mathcal{M}_1$ and $\mathcal{M}_2$, and wants to compare them based on *the posterior probability* of $\mathcal{M}_i$ given data Y, say $\Pi(\mathcal{M}_i \mid Y)$. Let $p(\mathcal{M}_i)$ be a prior distribution over the models. Then Bayesian model selection will be done via comparing the posterior model probability (Kass and Raftery, 1995) between models $\mathcal{M}_1$ and $\mathcal{M}_2$, which is given by:

$$\frac{\pi(\mathcal{M}_1 \mid Y)}{\pi(\mathcal{M}_2 \mid Y)} = \frac{p(\mathcal{M}_1)\int_{\Theta_1} p(\theta_1 \mid \mathcal{M}_1)p(Y \mid \mathcal{M}_1, \theta_1)d\theta_1}{p(\mathcal{M}_2)\int_{\Theta_2} p(\theta_2 \mid \mathcal{M}_2)p(Y \mid \mathcal{M}_2, \theta_2)d\theta_2} = \frac{p(\mathcal{M}_1)}{p(\mathcal{M}_2)} BF_{12} \tag{7.2}$$

where Θ_i and θ_i denote the parameter space and the parameter for the model $i = 1, 2$ and:

$$BF_{12} := \frac{\int_{\Theta_1} p(\theta_1 \mid \mathcal{M}_1)p(Y \mid \mathcal{M}_1, \theta_1)d\theta_1}{\int_{\Theta_2} p(\theta_2 \mid \mathcal{M}_2)p(Y \mid \mathcal{M}_2, \theta_2)d\theta_2}, \tag{7.3}$$

is called *the Bayes factor* (Kass and Raftery, 1995). From the definition above, it is clear to see of great importance of the evidence $\int_{\Theta_i} p(\theta_i \mid \mathcal{M}_i)p(Y \mid \mathcal{M}_i, \theta_i)d\theta_i$. In general, the posterior model probability has the consistency property. That is if we select the model via the posterior model probability (the Bayes factor), then, as $n \to \infty$, the probability that selects the true model goes to 1. See Chib and Kuffner (2016) for general treatment of the consistency of the posterior model probability, for instance.

Suppose that stochastic process $X := \{X_t; 0 \le t \le T\}$ is obtained by the solution to the following time-homogeneous stochastic differential equation (SDE):

$$dX_t = b_\zeta(X_t)dt + \sigma_\zeta(X_t)dB_t^H, \tag{7.4}$$

where $B^H = (B_t^H; 0 \leq t \leq T)$ is a *fractional Brownian motion* (fBm) which is a centered Gaussian process characterised by its *Hurst parameter* $H \in (0, 1)$, and has the covariance:

$$\text{Cov}\left(B_t^H, B_s^H\right) = \frac{1}{2}\left(\mid t \mid^{2H} + \mid s \mid^{2H} - \mid t - s \mid^{2H}\right). \tag{7.5}$$

For $H > \frac{1}{2}$, B^H exhibits long-range dependence in the sense that the autocorrelation function of fBm is not summable, and increments are positively correlated. In contrast, for $H < \frac{1}{2}$, increments of fBm are negatively correlated with rougher paths. That is, the its autocorrelation function is summable but decays slowly characterizing short-range dependence. When $H = \frac{1}{2}$, fBm becomes the well-known Brownian motion. Therefore, fBm can be thought of as a generalization of the Brownian motion allowing for memory in its increments.

Since the pioneering work of Mandelbrot and Van Ness (1968), owing to their rich structure regarding memory, models driven by fractional noise are being routinely used in such diverse disciplines, we refer to Kou (2008) for various applications of such models in science. In particular, due to the stylized facts suggesting the existence of the long memory property of the volatility process (Andersen and Bollerslev, 1997; Andersen et al., 2001; Ding et al., 1993; Cont, 2001; Lobato and Savin, 1998; Lobato and Velasco, 2000; Bollerslev and Jubinski, 1999), introducing a long memory in continuous time stochastic volatility models has been an object of research in finance and financial econometrics (Comte and Renault, 1996, 1998; Corsi, 2009; Chronopoulou and Viens, 2012; Guennoun et al., 2018; Comte et al., 2012). After decades, introducing a short-range dependence in continuous time stochastic volatility models also has been steadily gaining attention (Gatheral et al., 2018; Bayer et al., 2016; El Euch and Rosenbaum, 2018, 2019; Forde and Zhang, 2017; Fukasawa, 2017; Alòs et al., 2007).

In this study, as studied in Gloter and Hoffmann (2007); Xiao et al. (2011); Rao (2011), we assume that the process X can be observed at only discrete time instances $Y_0, Y_1, \ldots Y_n$, $0 \leq m \leq n$, $0 \leq t_1 < t_2 < \cdots t_n$ possibly with random errors parametrized by the parameter denoted by λ. Also we write the parameters $\theta := (\zeta, H, \lambda) \in \mathbb{R}^d$ and the likelihood of $Y := (Y_0, Y_1, \ldots Y_n)$ as:

$$p(Y \mid \theta) = \int p(Y \mid \theta, X) p(dX \mid \theta). \tag{7.6}$$

Throughout this study, we assume that the likelihood function of Y given B^H, say $p(Y \mid B^H)$, is analytically tractable but we cannot marginalize the model onto finite dimensions. Let $p(d\theta)$ denotes the prior distribution over parameters θ. Then Bayes' theorem yields:

$$p(\theta \mid Y) = \frac{p(\theta) p(Y \mid \theta)}{p(Y)}, \tag{7.7}$$

where:

$$p(Y) = \int_\Theta p(d\theta) p(Y \mid \theta), \tag{7.8}$$

The main objective of our study is to estimate the evidence in (7.8) for SDE models driven by fractional noise in (7.4) and make use of it for Bayesian model selection via the posterior model probability in (7.2). Model selection for models based on (7.4) is of particularly important since it will

enable us to shed light on the econometric debate on the short-range or long-range nature dependence in volatility (Cont, 2001). In particular we will develop a proper Markov chain Monte Carlo (MCMC) algorithm tailored to the structure of the models of interest. However, several challenging problems will arise in our setting.

The first problem is the intractability of the likelihood function in (7.6). To adders this problem, we adopt a data-augmentation approach from a Bayesian perspective. That is, we try to sample from the joint posterior density of (X, θ):

$$\Pi(X, \theta \mid Y) \propto p(Y \mid X, \theta) p(X \mid \theta) p(\theta). \tag{7.9}$$

However, sampling from (7.9) will end up with slow mixing due to the high correlation between B^H and H which yields the high correlation between X and H. When $H = \frac{1}{2}$, due to the Markovian property, decoupling such dependency is well documented (Kalogeropoulos et al., 2010; Roberts and Stramer, 2001; Golightly and Wilkinson, 2008), and also MCMC within particle methods can be straightforwardly applied (Andrieu et al., 2010; Chopin et al., 2013). Since we allow H to take values in $(0, 1)$, such techniques based on the Markovian property cannot be applied to our setting. In fact, when $H \neq \frac{1}{2}$, the parameters θ can be fully identified by a continuous path of X, and thus the joint posterior (7.7) will be degenerated, with $p(X \mid H)$ being a Dirac measure, see Rao (2011) for instance. In addition, most critically, estimating the evidence with an intractable likelihood in high-dimensional spaces is itself one of the most important challenges to computational methodology (Everitt et al., 2017; Lyne et al., 2015; Carlin and Chib, 1995; Chib, 1995; Gelfand and Smith, 1990). This problem will be of particular concern for our setting since a path of SDE is defined on an infinite-dimensional space with involving an intractable likelihood.

In this study, these challenging issues are addressed in order to develop a scalable, robust and adaptive MCMC algorithm. In particular, we resort to a class of Hamiltonian Monte Carlo (Duane et al., 1987) defined on an infinite-dimensional space within Sequential Monte Carlo (SMC) samplers studied in Del Moral et al. (2006). Compared with literature, the main contributions of the paper in this setting – and they do relate with overcoming the difficulties – will be as follows:

i) Following closely Beskos et al. (2015), we present a scalable method to simulate realizations of B^H with the $\mathcal{O}(N \log N)$ computational cost first studied in Davies and Harte (1987). This method also decouples the dependency between B^H and X, and thus gives rise to a-priori independent structure. Since this algorithm is well-defined for any $H \in (0, 1)$, we do not need to assume the particular range of values of H to do statistical inference, in contrast with the literature typically assuming $H \geq \frac{1}{2}$ such as Neuenkirch and Tindel (2014); Xiao and Yu (2019); Belfadli et al. (2011); Hu and Nualart (2010); Xiao et al. (2011). This ad hoc restriction is not really favourable since some empirical studies show that the Hurst parameter will be less than $\frac{1}{2}$ (Gatheral et al., 2018; Beskos et al., 2015; Cont, 2001; Bayer et al., 2016).

ii) The convergence property of our MCMC algorithm is mesh-free in the sense that its mixing time does not deteriorate as the dimension of the state space increases. This property is of particular important since, in practice, one has to discretize the continuous models in (7.4) via some numerical approximation methods, and this gives rise to an $N-$dimensional proxy of

the target which is theoretically defined on the corresponding infinite-dimensional space. What the mesh-free property says is that the convergence property will not deteriorate upon refinement of the approximation of the inherently infinite-dimensional diffusion paths of X by the finite $N-$dimensional proxy ones used in practice when applying the algorithms on a computer. Therefore, our algorithm is robust over the dimension of the models being considered.

iii) As we will see later, our algorithm involves some user-specified tuning parameters (hyper parameters) in advance, and the performance of the algorithm will be affected significantly by such parameters in practice. In this paper, we present a class of Hamiltonian Monte Carlo within sequential Monte Carlo samplers. This approach allows, given some initials, the tune parameters to be adaptively learned within the algorithm according to the information on the models and data by using the outputs generated by sequential Monte Carlo. In addition, as a result of the usage of sequential Monte Carlo samplers, we can construct an unbiased and consistent estimator of the evidence in (7.8), and thus the Bayes factor in (7.3) as well.

iv) We also develop MCMC for Bayesian model selection. Our algorithm can be considered as an exact approximation of the posterior model probability. Compared with other trans-dimensionl MCMC methods, our method can avoid finding a dimension-match transformation. Also, our model selection method could be well defined on the diffusion pathspace a consequence of the mesh-free property.

The rest of the paper is organised as follows. In subsection 7.2 we review basics of Bayesian model selection and its computational strategies. We then develop our methods for Bayesian model selection on high dimensional spaces in subsection 7.3. subsection 7.4 is devoted to providing our approach to overcome high correlation between B^H and H. We first develop efficient technique for simulating fractional Brownian motion, then we study joint inference for (X, θ) based on the advanced Hamiltonian Monte Carlo. The advanced Hamiltonian Monte Carlo involves some user specified tuning parameters. Thus we study how to make use of SMC samplers to learn adaptively tuning parameters in subsection 7.5, then conclude in subsection 7.6.

7.2 Basics of Bayesian model selection and computational strategies

Consider a countable set of parametric models, denoted by $\mathcal{M} = \{\mathcal{M}_k\}_{k\in\mathsf{K}}$, and let $Y = y_{0:n}$ be the data observed according to a likelihood function $\mathcal{L}(Y \mid \theta_k, \mathcal{M}_k)$ with corresponding parameter spaces $\theta_k \in \Theta_k$. In the framework of Bayesian model selection, one needs to specify a prior distribution must be specified over the parameter given the model and a model, we write $p(\theta_k \mid \mathcal{M}_k)$ and $p(\mathcal{M}_k)$ respectively. Then Bayes' theorem gives rise to the the posterior distribution of the parameters and the model given Y on the product space $\cup_{k\in\mathcal{K}} \{\mathcal{M}_k\} \times \Theta_k$:

$$\Pi(\theta_k, \mathcal{M}_k \mid Y) = \frac{\mathcal{L}(Y \mid \theta_k, \mathcal{M}_k) p(\theta_k \mid \mathcal{M}_k) p(\mathcal{M}_k)}{p(Y)}, \tag{7.10}$$

where:

$$p(Y) := \sum_{k \in \mathsf{K}} p(Y \mid \mathcal{M}_k) p(\mathcal{M}_k), \tag{7.11}$$

$$p(Y \mid \mathcal{M}_k) := \int \mathcal{L}(Y \mid \theta_k, \mathcal{M}_k) p\left(\theta_k \mid \mathcal{M}_k\right) d\theta_k. \tag{7.12}$$

In the context of Bayesian model selection, $\Pi\left(\theta_k, \mathcal{M}_k \mid Y\right)$ is often called the *full posterior*. Within the model, one obtains:

$$\Pi\left(\theta_k \mid Y, \mathcal{M}_k\right) = \frac{\mathcal{L}(Y \mid \theta_k, \mathcal{M}_k) p\left(\theta_k \mid \mathcal{M}_k\right)}{p(Y \mid \mathcal{M}_k)}, \tag{7.13}$$

where $p(Y \mid \mathcal{M}_k)$ is again called the *evidence* or the *marginal likelihood*. It turns out that *the posterior model probability* of the model $\mathcal{M}_k$ is given by θ_k-marginal of $\Pi(\theta_k, \mathcal{M}_k \mid Y)$:

$$\Pi\left(\mathcal{M}_k \mid Y\right) = \frac{\int \mathcal{L}(Y \mid \theta_k, \mathcal{M}_k) p\left(\theta_k \mid \mathcal{M}_k\right) p(\mathcal{M}_k) d\theta_k}{\sum_{l \in \mathsf{K}} p(\mathcal{M}_l) \int \mathcal{L}(Y \mid \theta_l, \mathcal{M}_l) p\left(\theta_l \mid \mathcal{M}_l\right) d\theta_l}. \tag{7.14}$$

Now assume that one has two models, say $\mathcal{M}_1$ and $\mathcal{M}_2$. Then the ratio of $\Pi\left(\mathcal{M}_1 \mid Y\right)$ and $\Pi\left(\mathcal{M}_2 \mid Y\right)$ is given by:

$$\begin{aligned} \frac{\Pi\left(\mathcal{M}_1 \mid Y\right)}{\Pi\left(\mathcal{M}_2 \mid Y\right)} &= \frac{p(\mathcal{M}_1) \int \mathcal{L}(Y \mid \theta_1, \mathcal{M}_1) p\left(\theta_1 \mid \mathcal{M}_1\right) d\theta_1}{p(\mathcal{M}_2) \int \mathcal{L}(Y \mid \theta_2, \mathcal{M}_2) p\left(\theta_2 \mid \mathcal{M}_2\right) d\theta_2}, \\ &= \frac{p(\mathcal{M}_1)}{p(\mathcal{M}_2)} B_{12}, \end{aligned} \tag{7.15}$$

where B_{12} is again called the *Bayes factor* given in (7.3). In other words, the posterior odds equal to the Bayes factor times the prior odds. Also notice that if one uses a uniform prior for $p(\mathcal{M}_k)$ for $k \in \mathsf{K}$ then (7.15) becomes just the Bayes factor so that it can be understood as the evidence provided by the data in favour of $\mathcal{M}_1$ against model $\mathcal{M}_2$ so that the Bayes factor is the principle tool for Bayesian model model selection. In particular, Kass and Raftery (1995); Jeffreys (1998) suggest the following interpretation of the the Bayes factor B_{12}.

$log_{10} B_{12}$	B_{12}	the evidence provided by the data in favour of $\mathcal{M}_1$ against model $\mathcal{M}_2$
0 to 2	1 to 3.2	Not worth more than a bare mention
2 to 6	3.2 to 10	Substantial
6 to 10	10 to 100	Strong
$10 > 0$	$100 > 0$	Decisive

Table 3: The interpretation of the Bayes factor in Kass and Raftery (1995); Jeffreys (1998).

Zhou et al. (2016) point out the following three fundamentally different approaches.

i) Calculate the posterior model probability $\Pi\left(\mathcal{M}_k \mid Y\right)$ directly.

ii) Calculate the evidence $p(Y \mid \mathcal{M}_k)$ of each model.

iii) Calculate pairwise evidence ratios, that is the Bayes factor $B_{k,l}$ for model $\mathcal{M}_k$ and $\mathcal{M}_l$ directly.

We then review several classical computational ways for the three approaches.

The first approach One of the straightforward ways could be using reversible jump MCMC Green (1995), see also Algorithm 2. Namely, one can explore the full posterior $\Pi(\theta_k, \mathcal{M}_k \mid Y) \propto \mathcal{L}(Y \mid \theta_k, \mathcal{M}_k) p(\theta_k \mid \mathcal{M}_k) p(\mathcal{M}_k)$ defined on the product space $\cup_{k\in\mathcal{K}} \{\mathcal{M}_k\} \times \Theta_k$ via reversible jump MCMC, and then integrate out w.r.t. θ_k. We again note that reversible jump MCMC requires a dimension matching transformation, and establishing such a transformation is itself a difficult problem.

The second approach One of the easiest ways to calculate $p(Y \mid \mathcal{M}_k)$, if one uses a uniform prior for $p(\mathcal{M}_k)$, might be calculating BIC of each model. As we studied, this strategy might work if the models satisfy the Laplace regular condition (Kass et al., 1991). Alternatively, let $\{\hat{\theta}^{(i)}\}_{i=1}^N$ be MCMC outputs targetting $\Pi(\theta_k \mid Y, \mathcal{M}_k)$. Since $p(Y \mid \mathcal{M}_k) := \int \mathcal{L}(Y \mid \theta_k, \mathcal{M}_k) p(\theta_k \mid \mathcal{M}_k) d\theta_k$, the harmonic mean (Newton and Raftery, 1994):

$$\hat{p}(Y \mid \mathcal{M}_k) := \left(\frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(Y \mid \hat{\theta}_k^{(i)}, \mathcal{M}_k) \right)^{-1}, \tag{7.16}$$

can be used as an approximation of $p(Y \mid \mathcal{M}_k)$ but it is well known that this estimator does not always have a finite variance (Kass and Raftery, 1995).

The third approach Consider a path of distributions such that $c = \gamma_\lambda(x)/\mathcal{Z}_\lambda$ where $\lambda \in [0,1]$ and $\mathcal{Z}_\lambda = \int \gamma_\lambda(x) dx$. Then we are interested in the logarithm of the ratio of their normalising constants, that is $r_{01} := \log(\mathcal{Z}_1/\mathcal{Z}_0)$. Clearly, this could be equal to the logarithm of the Bayes factor. In practice, one could set $\gamma_\lambda(x) = (1-\lambda)\gamma_0(x) + \lambda\gamma_1(x)$ for $\lambda \in [0,1]$. Under some assumptions, it can be shown that:

$$\nabla \log \mathcal{Z}_\lambda = \frac{\nabla \mathcal{Z}_\lambda}{\mathcal{Z}_\lambda} = \frac{\nabla \left[\int \gamma_\lambda(x) dx\right]}{\mathcal{Z}_\lambda} = \int \nabla \log \gamma_\lambda(x) \eta_\lambda(x) dx = \mathbb{E}_{\eta_\lambda} \left[\nabla \log \gamma_\lambda\right]. \tag{7.17}$$

Integrating (7.17), from 0 to 1 gives the log of the ratio of the normalising constants:

$$r_{01} = \log\left(\frac{\mathcal{Z}_1}{\mathcal{Z}_0}\right) = \int_0^1 \mathbb{E}_{\eta_\lambda} \left[\nabla \log \gamma_\lambda\right] d\lambda. \tag{7.18}$$

If one considers λ as a random variable with a uniform distribution on $[0,1]$, one could see (7.18) as the expectation of $\nabla \log \gamma_\lambda$ w.r.t. the joint distribution of (x, λ). Introducing a prior density $p(\lambda)$ for $\lambda \in [0,1]$ gives rise to the *path sampling identity*:

$$r_{01} = \int_0^1 \frac{\mathbb{E}_{\eta_\lambda} \left[\nabla \log \gamma_\lambda\right]}{p(\lambda)} p(\lambda) d\lambda. \tag{7.19}$$

Using (7.19), Gelman and Meng (1998) introduce the followings. First one discretises λ as L points on $[0,1]$, say $0 = \lambda^{(1)} < \ldots < \lambda^{(L)} = 1$. Then for each $l \in \{1, \cdots, L\}$ with $\lambda = \lambda^{(l)}$, obtain N_l MCMC

samples leaving $\eta_\lambda(x)$ invariant. Then one could estimate $\mathbb{E}_{\eta_\lambda}\left[\nabla \log \gamma_\lambda\right]$ by such MCMC outputs so that r_{01} could be obtained by numerical integration w.r.t. λ.

Zhou et al. (2016) argue that these three methods admit natural SMC samplers (Algorithm 12) based strategy, and each one has pros and cons. In this study, we will follow the first approach. As we mentioned, one of the main drawbacks of this approach is that requires a non-trivial transformation. To overcome this problem, we will use pseudo-marginal MCMC (see subsection 2.4) to approximate $\Pi\left(\theta_k, \mathcal{M}_k \mid Y\right)$ exactly. Also, we will tackle a problem arising from high-dimensional nature of diffusion models by using the advanced HMC (see subsection 2.7) within SMC sampler.

7.3 Estimating the evidence

We propose SMC sampler to estimate the evidence in subsubsection 7.3.1. A critical point is that one can construct an unbiased estimator of the evidence. Using such an estimator, we makes use of pseudo-marginal MCMC to approximate the posterior model probability. To overcome high-dimensional nature of the latent path driven by diffusion models, we then construct a MCMC kernel based on the advanced HMC whose mixing behaviour does not depend on the dimensionality.

7.3.1 Tempering and sequential Monte carlo sampler

First, we introduce *the tempered posterior distributions:*

$$\eta_n(dx) := \frac{\gamma_n(dx)}{\mathcal{Z}_n}, \tag{7.20}$$

where $\gamma_n(dx) := \mathcal{L}(x)^{\phi_n}\pi_0(dx)$, $\mathcal{Z}_n := \int \gamma_n(dx)$ and $0 = \phi_0 < \phi_1 < \cdots \phi_n < \cdots \phi_p = 1$, thus $\eta_p(dx) = \eta(dx)$ and $\eta_0(dx) = \pi_0(dx)$. The sequence $\{\phi_n\}_{n=0}^p$ is commonly called inverse temperature. Note that the index n is auxiliary. As noted by Chopin (2002); Del Moral et al. (2006); Neal (2001); Zhou et al. (2016), such a tempering approach might provide potential stability and reduction in computational complexity in a high-dimensional setting. The sequence of the tempered distributions can be also understood as a sequence of the bridging distributions in the sense that they gradually evolve from the tractable prior $\pi_0(dx) = \eta_0(dx)$ to the complex the posterior $\eta(dx) = \eta_p(dx)$. As a consequence, now we have a sequence of probability distributions $\{\hat{\eta}_n(dx)\}_{n=0}^p$ which are defined on a common measurable space. SMC cannot be applied directly to such a sequence of distributions since it is available for distributions whose dimension is increasing over time index. See Doucet and Johansen (2009) and references therein, and section 3 for a more in-depth treatment of SMC.

As we studied in subsection 3.3, we make use of SMC sampler (Algorithm 12). As before, define a sequence of distributions defined on product spaces $(E^p, \mathcal{E}^p) := (\prod_{n=0}^p E^n, \mathcal{E}^{\otimes p})$:

$$\tilde{\eta}_n(x_{1:n}) := \eta_n(x_n) \prod_{k=1}^{n-1} \mathcal{B}_k(x_{k+1}, x_k), \tag{7.21}$$

where $\mathcal{B}_k$ is a transition density from E^{k+1} to E^k. Also let $\{\mathcal{K}_k\}_{k=1}^{n-1} : E \times \mathcal{E} \to [0,1]$ be $\eta_k(dx)-$reversible MCMC kernels which are chosen to satisfy that $\mathcal{B}_{k-1} \otimes \gamma_k$ is absolutely continuous w.r.t. $\gamma_{k-1} \otimes \mathcal{K}_k$ for any k. Then it is given by $\mathcal{B}_k(x_{k+1}, x_k) = \frac{\eta_k(dx_{k-1})\mathcal{K}_k(\theta_{k-1}, \theta_k)}{\eta_k(dx_k)}$ so that we have unnormalised

incremental weight:

$$w_k := \frac{\mathcal{B}_{k-1}(x_k, x_{k-1})}{\mathcal{K}_k(x_{k-1}, x_k)} \frac{\gamma_k(x_{k-1})}{\gamma_{k-1}(x_{k-1})} = \mathcal{L}(x_{k-1})^{(\phi_k - \phi_{k-1})}. \tag{7.22}$$

Critically, notice that:

$$\int w_k \eta_{k-1}(x_{k-1}) \mathcal{K}_k(x_{k-1}, x_k) dx_{k-1:k} = \int \mathcal{B}_{k-1}(x_n, x_{k-1}) \frac{\gamma_k(x_k)}{\mathcal{Z}_{k-1}} dx_{k-1:k} = \frac{\mathcal{Z}_k}{\mathcal{Z}_{k-1}},$$

and thus, this gives rise to the estimate of the ratio of normalising constant $\frac{\mathcal{Z}_k}{\mathcal{Z}_{k-1}}$ as follows:

$$\begin{cases} \widehat{\frac{\mathcal{Z}_k}{\mathcal{Z}_{k-1}}} & = \sum_{i=1}^N W_{k-1}^{(i)} w_k^{(i)}, \\ W_k^{(i)} & := \frac{w_k^{(i)}}{\sum_{j=1}^N w_k^{(j)}}. \end{cases} \tag{7.23}$$

Thus, by-product of (7.23), we can obtain the estimate of $\mathcal{Z}_n$: for $n \geq 1$:

$$\widehat{\mathcal{Z}}_n = \widehat{\mathcal{Z}}_0 \prod_{k=1}^n \left(\sum_{i=1}^N W_{k-1}^{(i)} w_k^{(i)} \right). \tag{7.24}$$

For the sake of completeness, we include the following proposition.

Proposition 41.

i) *For any* $N, n \geq 0$, $\mathbb{E}\left[\widehat{\mathcal{Z}}_n\right] = \mathcal{Z}_n$ *holds.*

ii) *For any* $N, n \geq 0$, $\widehat{\mathcal{Z}}_n \geq 0$.

iii) *Assume that for any* $x \in E$, *there exists* $c \in (0,1)$ *such that* $c \leq \mathcal{L}(x)^{\Delta\phi_n} \leq c^{-1}$ *where* $\Delta\phi_n := \phi_n - \phi_{n-1}$ *for* $n \geq 1$. *Then we have that for any* $n \geq 0$, $\sqrt{\mathbb{E}\left[\left(\widehat{\mathcal{Z}}_n - \mathcal{Z}_n\right)^2\right]} \to 0$ *as* $N \to \infty$.

Proof. Set the potential function as $G(x_{n-1}, x_n) := \frac{\gamma_n(x_{n-1})}{\gamma_{n-1}(x_{n-1})}$, and define $\mathbb{Q}(dx_{1:n}) := \eta_n(dx_n) \prod_{k=1}^{n-1} \mathcal{B}_k(x_{k+1}, x_k)$. Then, as we studied in Example 13, the pair $(G(x_{n-1}, x_n), \mathcal{K}_n)$ recovers the Feynman-Kac path measure $\mathbb{Q}(dx_{1:n})$. Indeed, it can be shown that, for $f \in \mathcal{B}_b(E)$, $\int f(x_n) \prod_{p=1}^{n-1} G(x_{p-1}, x_p) \mathcal{K}_p(x_{p-1}, x_p) dx_{0:p} = \int f(x_n) \gamma_n(x_n) dx_n =: \gamma_n(f)$, and thus $\eta_n(f) = \frac{\gamma_n(f)}{\gamma_n(1)}$ holds. Therefore, the claims follow from Theorem 17 and Theorem 18 respectively. □

7.3.2 MCMC for the Bayesian model selection

Suppose that now one has a countable set of candidate models, denoted by $\mathcal{M} := \{\mathcal{M}_k\}_{k \in \mathsf{K}}$ with corresponding parameter spaces $\theta_k \in \Theta_k \subseteq \mathbb{R}^{d_k}$. Then we have the tempered full posterior density on the space $\cup_{k \in \mathcal{K}} \{\mathcal{M}_k\} \times \Theta_k$ such that:

$$\eta_n(\theta_k, \mathcal{M}_k \mid Y) = \frac{\mathcal{L}(Y \mid \theta_k, \mathcal{M}_k)^{\phi_n} \pi_0(\theta_k) p(\mathcal{M}_k)}{\sum_{k \in \mathcal{K}} \mathcal{Z}_n(Y \mid \mathcal{M}_k) p(\mathcal{M}_k)}, \tag{7.25}$$

where again Y denotes data, $p(\mathcal{M}_k)$ is a discrete prior distribution over $\mathcal{M}_k$ and:

$$\mathcal{Z}_n(Y \mid \mathcal{M}_k) := \int_{\Theta_k} \mathcal{L}(Y \mid \theta_k, \mathcal{M}_k)^{\phi_n} \pi_0(\theta_k) d\theta_k. \tag{7.26}$$

From (7.25), again, it is clear to see that the posterior probability of the model $\mathcal{M}_k$ is given by θ_k-marginal of $\eta(\theta_k, \mathcal{M}_k \mid Y)$:

$$\eta_n(\mathcal{M}_k \mid Y) = \frac{p(\mathcal{M}_k) \int_{\Theta_k} \mathcal{L}(Y \mid \theta_k, \mathcal{M}_k)^{\phi_n} \pi_0(\theta_k) d\theta_k}{\sum_{l \in \mathcal{K}} p(\mathcal{M}_l) \mathcal{Z}_n(Y \mid \mathcal{M}_l)} = \frac{p(\mathcal{M}_k) \mathcal{Z}_n(Y \mid \mathcal{M}_k)}{\sum_{l \in \mathcal{K}} p(\mathcal{M}_l) \mathcal{Z}_n(Y \mid \mathcal{M}_l)}, \tag{7.27}$$

which has been considered as the core principle within the Bayesian model comparison problem, see e.g. Chipman et al. (2001); Robert (2007); Kass and Raftery (1995). In order to carry out Bayesian model selection, we want to establish a MCMC for (7.27) on the space $\cup_{k \in \mathcal{K}} \{\mathcal{M}_k\} \times \Theta_k$. Let $q_{\mathcal{M}}(\mathcal{M}_{k'} \mid \mathcal{M}_k)$ be a some known proposal. Then the ideal MCMC can be algorithmically described as:

i) Propose a candidate $\mathcal{M}_{k'}$ via $q_{\mathcal{M}}(\mathcal{M}_{k'} \mid \mathcal{M}_k)$

ii) Accept $\mathcal{M}_{k'}$ w.p. $\min\left\{1, \frac{q_{\mathcal{M}}(\mathcal{M}_k \mid \mathcal{M}_{k'}) \eta(\mathcal{M}_{k'} \mid Y)}{q_{\mathcal{M}}(\mathcal{M}_{k'} \mid \mathcal{M}_k) \eta(\mathcal{M}_k \mid Y)}\right\} = \min\left\{1, \frac{q_{\mathcal{M}}(\mathcal{M}_k \mid \mathcal{M}_{k'}) p(\mathcal{M}_{k'}) \mathcal{Z}(Y \mid \mathcal{M}_{k'})}{q_{\mathcal{M}}(\mathcal{M}_{k'} \mid \mathcal{M}_k) p(\mathcal{M}_k) \mathcal{Z}(Y \mid \mathcal{M}_k)}\right\}$.

The reason why above algorithm is the ideal one is that, as we mentioned, intractability of the model evidence $\mathcal{Z}(Y \mid \mathcal{M}_k)$ in (7.26). Therefore, Bayesian model selection directly based on the posterior probability (7.27) cannot be implemented in practice. To adders this problem, we propose the following procedure which is an application of pseudo-marginal MCMC (Andrieu and Roberts, 2009), and a similar idea can be found in Karagiannis and Andrieu (2013).

The main idea is that whilst $\mathcal{Z}(Y \mid \mathcal{M}_k)$ is intractable, we can construct an estimator of it via Algorithm 12, that is, (7.24) given $\mathcal{M}_k$. This estimator is indeed unbiased due to Proposition 41. Then this leads us to replace the target $\eta(\mathcal{M}_k \mid Y)$ by the extended exact approximation based on the unbiased estimator, which marginally admits the target. We note that both the ideal method and the one which we will introduce can be considered as a Gibbs type method in the sense that we first run a simulation only on Θ_k given $\mathcal{M}_k$ instead of sampling jointly from (7.25) on the space $\cup_{k \in \mathcal{K}} \{\mathcal{M}_k\} \times \Theta_k$.

To sample from (7.27), we first run Algorithm 12 until $n = p$, and write all random variables generated by Algorithm 12 as $\mathcal{U}_k \in \mathsf{U}_k$ so that now we can construct the extended approximate target $\widehat{\eta}(\mathcal{M}_k, \mathcal{U}_k \mid Y)$ on the extended product space $\mathcal{M}_k \times \mathsf{U}_k$. To be precise, we have that:

$$\hat{\eta}(\mathcal{M}_k, \mathcal{U}_k \mid Y) \propto p(\mathcal{U}_k) p(\mathcal{M}_k) \hat{\mathcal{L}}(Y \mid \mathcal{M}_k, \mathcal{U}_k), \tag{7.28}$$

where $\hat{\mathcal{L}}(Y \mid \mathcal{M}_k, \mathcal{U}_k)$ is given in (7.24) at auxiliary time $n = p$, $p(\mathcal{U}_k)$ denotes the density of $\mathcal{U}_k$, and we have assumed that $\mathcal{M}_k$ and $\mathcal{U}_k$ are a-priori independent for any $k \in \mathsf{K}$. Note that here we use $\mathcal{U}_k$ to emphasise the dependency of a particle estimate of $\mathcal{Z}(Y \mid \mathcal{M}_k)$ on $\mathcal{U}_k$. In this case, a joint proposal density of $(\mathcal{M}_k, \mathcal{U}_k)$ is given by:

$$q(\mathcal{M}_{k'}, \mathcal{U}_{k'} \mid \mathcal{M}_k, \mathcal{U}_k) = q_{\mathcal{M}}(\mathcal{M}_{k'} \mid \mathcal{M}_k, \mathcal{U}_k) q_{\mathcal{U}}(\mathcal{U}_{k'} \mid \mathcal{U}_k), \tag{7.29}$$

which is the product of the two proposals that we set as:

$$q_{\mathcal{M}}(\mathcal{M}_{k'} \mid \mathcal{M}_k, \mathcal{U}_k) = q_{\mathcal{M}}(\mathcal{M}_{k'} \mid \mathcal{M}_k),\ q_{\mathcal{U}}(\mathcal{U}_{k'} \mid \mathcal{U}_k) = p(\mathcal{U}_{k'}). \tag{7.30}$$

Finally, we accept $\mathcal{M}_{k'}$ and $\mathcal{U}_{k'}$ w.p.

$$\begin{aligned} \alpha\left((\mathcal{M}_k, \mathcal{U}_k), (\mathcal{M}_{k'}, \mathcal{U}_{k'})\right) &:= \min\left\{1, \frac{q(\mathcal{M}_k, \mathcal{U}_k \mid \mathcal{M}_{k'}, \mathcal{U}_{k'})\hat{\eta}(\mathcal{M}_{k'}, \mathcal{U}_{k'} \mid Y)}{q(\mathcal{M}_{k'}, \mathcal{U}_{k'} \mid \mathcal{M}_k, \mathcal{U}_k)\hat{\eta}(\mathcal{M}_k, \mathcal{U}_k \mid Y)}\right\}, \\ &= \min\left\{1, \frac{q_{\mathcal{M}}(\mathcal{M}_k \mid \mathcal{M}_{k'})\hat{\mathcal{L}}(Y \mid \mathcal{M}_{k'}, \mathcal{U}_{k'})p(\mathcal{M}_{k'})}{q_{\mathcal{M}}(\mathcal{M}_{k'} \mid \mathcal{M}_k)\hat{\mathcal{L}}(Y \mid \mathcal{M}_k, \mathcal{U}_k)p(\mathcal{M}_k)}\right\}, \end{aligned} \tag{7.31}$$

then these developments lead us to have the followings.

Proposition 42. *Given $\mathcal{M}_k$ and $\mathcal{U}_k$, consider the following procedure.*

i) Propose a move from the model $\mathcal{M}_k$ to the model $\mathcal{M}_{k'}$ via proposal density $q_{\mathcal{M}}(\mathcal{M}_{k'} \mid \mathcal{M}_k)$.

ii) Given $\mathcal{M}_{k'}$, run Algorithm 12 untill auxiliary time $n = p$, and obtain $\hat{\mathcal{L}}(Y \mid \mathcal{M}_{k'}, \mathcal{U}_{k'})$.

iii) Accept $(\mathcal{M}_{k'}, \mathcal{U}_{k'})$ according to (7.31).

Then we have that:

i) $\int \hat{\eta}(\mathcal{M}_k, \mathcal{U}_k \mid Y) d\mathcal{U}_k = \eta(\mathcal{M}_k \mid Y)$ *for any* $k \in \mathsf{K}$.

ii) *The Markov chain* $(\mathcal{M}_k, \mathcal{U}_k) \mapsto (\mathcal{M}_{k'}, \mathcal{U}_{k'})$ *induced by above has* $\hat{\eta}(\mathcal{M}_{k'}, \mathcal{U}_{k'} \mid Y)$ *as an invariant distribution for any* $k' \in \mathsf{K}$.

iii) *The Markov chain* $(\mathcal{M}_k, \mathcal{U}_k) \mapsto (\mathcal{M}_{k'}, \mathcal{U}_{k'})$ *induced by above is uniformly ergodic for any* $k' \in \mathsf{K}$.

Proof. Since $\hat{\mathcal{L}}(Y \mid \mathcal{M}_k, \mathcal{U}_k)$ is unbiased (Proposition 41), we have that:

$$\int \hat{\mathcal{L}}(Y \mid \mathcal{M}_k, \mathcal{U}_k) p(\mathcal{U}_k) d\mathcal{U}_k = \mathcal{L}(Y \mid \mathcal{M}_k).$$

As a consequence, we obtain:

$$p(\mathcal{M}_k) \int \hat{\mathcal{L}}(Y \mid \mathcal{M}_k, \mathcal{U}_k) p(\mathcal{U}_k) d\mathcal{U}_k \propto \eta(\mathcal{M}_k \mid Y),$$

as required. Also, we have that:

$$\begin{aligned} \hat{\mathcal{L}}(Y \mid \mathcal{M}_k, \mathcal{U}_k)) q_{\mathcal{M}}(\mathcal{M}_{k'} \mid \mathcal{M}_k) p(\mathcal{U}_{k'}) \alpha\left((\mathcal{M}_k, \mathcal{U}_k), (\mathcal{M}_{k'}, \mathcal{U}_{k'})\right) \propto p(\mathcal{U}_k) p(\mathcal{U}_{k'}) \times \\ \min\left\{p(\mathcal{M}_k)\hat{\mathcal{L}}(Y \mid \mathcal{M}_k, \mathcal{U}_k) q_{\mathcal{M}}(\mathcal{M}_{k'} \mid \mathcal{M}_k), p(\mathcal{M}_{k'})\hat{\mathcal{L}}(Y \mid \mathcal{M}_{k'}, \mathcal{U}_{k'}) q_{\mathcal{M}}(\mathcal{M}_k \mid \mathcal{M}_{k'})\right\} \end{aligned}$$

which is clearly symmetric w.r.t. $(\mathcal{M}_{k'}, \mathcal{U}_{k'}) \leftrightarrow (\mathcal{M}_k, \mathcal{U}_k)$ so that the detailed balance condition w.r.t. $\hat{\eta}(\mathcal{M}_k, \mathcal{U}_k \mid Y)$ holds. The final claim of the proposition follows from the fact that the ideal MCMC is obviously uniformly ergodic since its state space K is finite, and thus Andrieu and Roberts (2009, Theorem 8) can be applied. □

Remark 21. *In the implementation, the result* $\int \hat{\eta}(\mathcal{M}_k, \mathcal{U}_k \mid Y) d\mathcal{U}_k = \eta(\mathcal{M}_k \mid Y)$ *means that* $\mathcal{U}_k$ *can be simply disregarded so that we will not store them in the subsequent implementations but only save the scalar value of estimate* $\hat{\mathcal{L}}(Y \mid \mathcal{M}_k, \mathcal{U}_k)$.

Above method can be considered as an exact approximation of the target $\eta(\mathcal{M}_k \mid Y)$, and therefore our approach has several advantages. First of all, we can avoid potential problems arising from trans-dimensional MCMC methods which aim to sample directly from the full posterior density (7.25). For instance, reversible jump MCMC (Green, 1995) has been routinely used in the context of Bayesian model selection. The reversible jump MCMC involves so-called dimension matching mapping due to joint inference on on the space $\cup_{k \in \mathcal{K}} \{\mathcal{M}_k\} \times \Theta_k$, which is rather hard to be established in general (Brooks et al., 2003), also it depends strongly on the models being considered. See also subsection 2.3. This is particularly problematic in a high-dimensional setting, and constructing such dimension matching mapping is less clear. Our approach does not require such a dimension matching mapping since it exactly approximates the idealised algorithm, and thus might outperform reversible jump MCMC whose mixing speed is typically poor unless one can find a good choice of the dimension matching mapping (Zhou et al., 2016; Karagiannis and Andrieu, 2013). Also, our method is an all in one approach (Zhou et al., 2016) in the sense that model selection and parameter estimates are simultaneously done. Therefore, one does not need to run a separate simulation for each model so that it might reduce the computational cost.

7.3.3 Constructing MCMC kernels on high dimensional spaces

Performance of SMC sampler (Algorithm 12) critically depends on a choice of MCMC kernels $\{\mathcal{K}_k\}_{k=1}^{n-1}$. Assume that $(E, \mathcal{E}) = (\mathbb{R}^N, \mathcal{B}(\mathbb{R}^N))$. Then the standard Hamiltonian Monte Carlo (HMC), first presented in Duane et al. (1987), involves extending the state space via introducing the auxiliary random variable called *velocity* $v \in \mathbb{R}^N$. Let $x \in \mathbb{R}^N$. x can be thought of as *location* and the *phase space* is defined as the product space of x and v, that is $\mathbb{R}^{2N}$. We refer to Neal (2011); Bou-Rabee and Sanz-Serna (2018) for a general reference of the HMC. See also subsection 2.6.

Unfortunately, HMC provides an inappropriate proposal $x^\star$ for increasing N. Indeed, the results in Beskos et al. (2013b) imply that one has to decrease the step size ϵ in a numerical integrator to $\mathcal{O}\left(N^{-1/4}\right)$ in order to control the acceptance probability for increasing N. Otherwise, critically, the acceptance probability will degenerate into 0 as $N \to \infty$ given ϵ. This difficultiy motivates us to use the *advanced Hamiltonian Monte Carlo* (Beskos et al., 2011, 2013a) which successfully avoids this degeneracy problem by employing a modified leapfrog integrator which yields better performance in high dimensions. That is, as we will study later, the advanced HMC has the important *mesh-free property* in the sense that the speed of mixing of the Markov chain does not deteriorate as $N \to \infty$. In other words, we have the fixed leapfrog step size $\epsilon = \mathcal{O}(1)$ even when $N \to \infty$. Clearly, this mesh-free property is particularly desirable for our setting.

To derive the algorithm, we first need to make the following assumption on the tempered target distribution $\eta_n(dx)$ in (7.20). First we define the *potential function*:

$$\Phi_n(x) := -\log \mathcal{L}(x)^{\phi_n} - \log \pi_0(x). \tag{7.32}$$

Assumption 18. *Let Π_0 be a centred Gaussian distiribution $\mathcal{N}(0, \mathcal{C})$. Then $\eta_n(dx)$ can be expressed as change of measure from Gaussian law:*

$$\frac{d\eta_n}{d\Pi_0}(x) \propto \exp\left(-\Phi_n(x)\right).$$

Cleary, Assumption 18 implies that now the density of $\eta_n(dx)$ is given by:

$$\eta_n(x) \propto \exp\left(-\Phi_n(x) - \frac{1}{2}\left\langle x, Lx\right\rangle\right), \tag{7.33}$$

where $L := \mathcal{C}^{-1}$. Then, the corresponding separating the Hamiltonian function on $\mathbb{R}^N \times \mathbb{R}^N$ is given by:

$$\mathsf{H}(x, v) = \Phi_n(x) + \frac{1}{2}\left\langle x, Lx\right\rangle + \frac{1}{2}\left\langle v, Mv\right\rangle, \tag{7.34}$$

where $\frac{1}{2}v^\top Mv$ is called the *kinetic* function with a user-specified positive-definite *mass matrix* M. Following Beskos et al. (2011), we set $M = L$ so that $\mathcal{C}^{-1} = M = L$. Then (7.34) can be decomposed into the two parts:

$$\mathsf{H}(x, v) = \mathsf{H}_1 + \mathsf{H}_2,\ \mathsf{H}_1 := \Phi(x)\ \mathsf{H}_2 := +\frac{1}{2}\left\langle x, Mx\right\rangle + \frac{1}{2}\left\langle v, Mv\right\rangle \tag{7.35}$$

It turns out that the corresponding Hamiltonian equations also can be split into:

$$\begin{cases} \frac{dx}{d\tau} = M^{-1}\frac{\partial \mathsf{H}_1}{\partial v} = 0, & \frac{dv}{d\tau} = -\frac{\partial \mathsf{H}_1}{\partial x} = -M^{-1}\nabla\Phi(x), \\ \frac{dx}{d\tau} = M^{-1}\frac{\partial \mathsf{H}_2}{\partial v} = v, & \frac{dv}{d\tau} = -\frac{\partial \mathsf{H}_2}{\partial x} = -x. \end{cases} \tag{7.36}$$

Notice that the system of the ordinary differential equations (7.36) can be solved analytically. That is, we can define the solution operators $\tilde{\Xi}^1_\tau, \tilde{\Xi}^2_\tau : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}^{2N}$ analytically:

$$\tilde{\Xi}^1_\tau := \left(x, v - \tau M^{-1}\nabla\Phi(x)\right), \tag{7.37}$$

$$\tilde{\Xi}^2_\tau := \left(\cos(\tau)x + \sin(\tau)v, -\sin(\tau)x + \cos(\tau)v\right). \tag{7.38}$$

respectively. Then, we can construct a discrete dynamical system corresponding to (7.36) as follows:

$$\Psi^{(1)}_\epsilon := \tilde{\Xi}^1_{\epsilon/2} \circ \tilde{\Xi}^2_{\epsilon^\star} \circ \tilde{\Xi}^1_{\epsilon/2}, \tag{7.39}$$

for $k \in \{0, \cdots, L-1\}$. It can be shown that Ψ_ϵ is indeed volume preserving and time reversible. Moreover, if we set $\cos(\epsilon^\star) = \frac{1-\epsilon^2/4}{1+\epsilon^2/4}$, then the integrator can be equivalently expressed as:

$$\begin{aligned} v_{k+1/2} &= v_k - \frac{\epsilon}{2}\nabla\Phi(x_k), \\ x_{k+1} &= \xi_k + \epsilon M^{-1} v_{k+1/2}, \\ v_{k+1} &= v_{k+1/2} - \frac{\epsilon}{2}\nabla\Phi(x_k), \end{aligned} \tag{7.40}$$

so that the exact flow induced by (7.35) can be approximated by the following L iterative steps:

$$\Psi_\epsilon^{\circ(k+1)} := \Psi_\epsilon^{\circ(k)} \circ \Psi_\epsilon^{\circ(1)}, \tag{7.41}$$

$k \in \{0, \cdots, L-1\}$, that is, we have that $x_L = \mathfrak{proj}^x \circ \Psi_\epsilon^{\circ(L)}(x_0, v_0)$. Then the acceptance probability has the same expression as for the standard HMC but with different notations, that is:

$$\alpha_{\mathsf{H}} := \min\left\{1, \exp\left(\mathsf{H}(x_0, v_{n0}) - \mathsf{H}(x^\star, -v^\star)\right)\right\}, \tag{7.42}$$

see Algorithm 6. Also recall that Markov kernel induced by the advanced HMC leaves the target $\eta_n(x)$ invariant under Assumption 18, see Proposition 22.

7.4 Simulating fractional Brownian motion and joint inference

Following Beskos et al. (2015) closely, we study an efficient and exact method to simulate increments of the fractional Brownian motion. Critically, this method gives rise to the mapping which transforms $2N$ independent standard Gaussian variables into fractional Gaussian noise at N discrete time instances. Using such a mapping, we then develop the advanced HMC for joint inference for (X, θ).

7.4.1 The Davies and Harte method and Decoupling dependency

We first study a method to decouple the dependency. As we mentioned, in practice, the algorithm has to be done with driving noise $\left\{B_t^H; 0 \le t \le T\right\}$ on a grid of discrete times. Define $t_j = jT/N$ for $j = 0, 1, \cdots N$ with the mesh size $\delta := \frac{T}{N}$. Typically, we set $T = 1$. We write $t_j = j$ for the sake of simplicity. Since fBm is self-similar and its increments are stationary (Rao, 2011), it is enough to simulate realisations $\{B_i^H\}_{i=1}^N$ and multiply them by δ^H. To simulate the such values, define the increments:

$$G_j^H := B_j^H - B_{j-1}^H,\ 1 \le j \le N, \tag{7.43}$$

with $G_1^H = B_1^H$. Then it turns out that, due to (7.5), the random variables $\{G_j^H\}_{j=1}^N$ are stationary standard Gaussian variables with the (auto) covariance, for $j \ge 1$:

$$\gamma(j) := \mathbb{E}\left[G_1^H G_{j+1}^H\right] = \frac{1}{2}\left((j+1)^{2H} + (j-1)^{2H} - 2j^{2H}\right), \tag{7.44}$$

this is so-called fractional Gaussian noise (fGn). Then a direct way to generate realisations of the fBm might be making use of the Cholesky decomposition. We begin with this approach to facilitate our study. Let Σ denote the $N \times N$ covariance matrix of the random vector $G^H := \left\{G_j^H; 1 \le j \le N\right\}$,

that is:

$$\Sigma := \begin{bmatrix} 1 & \gamma(1) & \gamma(2) & \cdots & \gamma(N-2) & \gamma(N-1) \\ \gamma(1) & 1 & \gamma(1) & \cdots & \gamma(N-3) & \gamma(N-2) \\ \gamma(2) & \gamma(1) & 1 & \cdots & \gamma(N-4) & \gamma(N-3) \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \gamma(N-2) & \gamma(N-3) & \gamma(N-4) & \cdots & 1 & \gamma(1) \\ \gamma(N-1) & \gamma(N-2) & \gamma(N-3) & \cdots & \gamma(1) & 1 \end{bmatrix}. \tag{7.45}$$

Then Cholesky method gives rise to the decomposition such that $\Sigma = \Gamma\Gamma^\top$. Given $u \sim \mathcal{N}(0, I_N)$ where I_N denotes the $N \times N$ identity matrix, one can easily simulate such realisations by setting $v = u\Gamma$. However, performing the Cholesky decomposition involves the computational cost $\mathcal{O}(N^3)$, and thus it is not feasible, in practice.

Following Wood and Chan (1994), we resort to the Davies-Harte method (Davies and Harte, 1987) to simulate fBm exactly. The Davies-Harte method relies upon the *Toeplitz* (or Diagonal-Constant) structure of Σ and the fast Fourier transform (FFT). Critically, the computational cost of the Davies-Harte methods is $\mathcal{O}(N \log N)$ and, to the best of our knowledge, this method is the fastest exact algorithm to simulate increments of B^H. Assume that $N = 2^g$ for $g \in \mathbb{N}$. In addition to Σ, we introduce the auxiliary matrix as follows:

$$\Sigma^f := \begin{bmatrix} 0 & \gamma(N-1) & \cdots & \gamma(2) & \gamma(1) \\ \gamma(N-1) & 0 & \cdots & \gamma(3) & \gamma(2) \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ \gamma(1) & \gamma(2) & \cdots & \gamma(N-1) & 0 \end{bmatrix}. \tag{7.46}$$

Then, critically, one can embed the Toeplitz matrix Σ into the following $2N \times 2N = 2^{g+1} \times 2^{g+1}$ *circular* matrix:

$$\begin{aligned} C &:= \begin{bmatrix} 1 & \gamma(1) & \cdots & \gamma(N-1) & 0 & \gamma(N-1) & \gamma(N-2) & \cdots & \gamma(2) & \gamma(1) \\ \gamma(1) & 1 & \cdots & \gamma(N-2) & \gamma(N-1) & 0 & \gamma(N-1) & \cdots & \gamma(3) & \gamma(2) \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \gamma(N-1) & \gamma(N-2) & \cdots & 1 & \gamma(1) & \gamma(2) & \gamma(3) & \cdots & \gamma(N-1) & 0 \\ 0 & \gamma(N-1) & \cdots & \gamma(1) & 1 & \gamma(1) & \gamma(2) & \cdots & \gamma(N-2) & \gamma(N-1) \\ \gamma(N-1) & 0 & \cdots & \gamma(2) & \gamma(1) & 1 & \gamma(1) & \cdots & \gamma(N-3) & \gamma(N-2) \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \gamma(1) & \gamma(2) & \cdots & 0 & \gamma(N-1) & \gamma(N-2) & \gamma(N-3) & \cdots & \gamma(1) & 1 \end{bmatrix}, \\ &= \begin{bmatrix} \Sigma & \Sigma^f \\ \Sigma^f & \Sigma \end{bmatrix}. \end{aligned} \tag{7.47}$$

Since C is circular, this allows a tractable eigen-expansion based on the FFT for C. Let P denotes the $2N \times 2N$ unitary matrix whose elements are given by $P_{jk} = (2N)^{-1/2} \exp(-\pi ijk/N)$ for $j, k = 0, \cdots 2N-1$ where $i^2 = -1$. Let P^* be the Hermitian transpose of P. We also define the diagonal

matrix $\Lambda_H := diag\left(\lambda_0, \lambda_1 \cdots, \lambda_{2N-1}\right)$ with the following eigenvalues, for $j, k = 0, \cdots 2N-1$:

$$\lambda_k := \sum_{k=0}^{2N-1} c_j \exp\left(-\pi i j k / N\right), \tag{7.48}$$

where $\{c_j\}$ are the components of the first row of the circular matrix C. Then C can be decomposed into:

$$C = P \Lambda_H P^*. \tag{7.49}$$

Using the FFT, one can calculate the components of Λ_H in $\mathcal{O}(N \log N)$ operations, see Golub and Van Loan (2012) for instance. Given this decomposition, the square root of C can be obtained in $\mathcal{O}(N)$ operations:

$$S := C^{1/2} = P \Lambda_H^{1/2} P^*, \tag{7.50}$$

which satisfies $SS^* = SS^\top = C$. Then this observation immediately gives rise to the followings. First one simulates Z from $\mathcal{N}(0, I_{2N})$ and obtain S. Then, given Z and S, the first N values of $SZ = P\Lambda_H^{1/2} P^* Z$ exactly provide the required fBm sample. Moreover, proposition 3 of Wood and Chan (1994) implies that the $\mathcal{O}(N \log N)$ computational costs of obtaining $P^* Z$ can be further reduce to the constant $O(N)$ computational costs. Again, given $Z \sim N(0, I_{2N})$, it can be shown that computing $P^* Z$ is equivalent to finding the matrix Q such that:

$$P \Lambda_H^{1/2} Q Z, \tag{7.51}$$

where Q is the $2N \times 2N$ sparse matrix such that:

$$Q := \begin{bmatrix} Q^{11} & Q^{12} \\ Q^{21} & Q^{22} \end{bmatrix}, \tag{7.52}$$

which has the $N \times N$ sub-matrices defined as the following way:

- $Q^{11} := diag\left\{1, 1/\sqrt{2}, 1/\sqrt{2}, \cdots, 1/\sqrt{2}\right\}$,
- $Q^{12} := \{q_{i,j}\}$, with $q_{i,i-1} = 1/\sqrt{2}$ for $1 \leq i \leq N-1$ and $q_{i,j} = 0$ otherwise,
- $Q^{21} := \{q_{i,j}\}$, with $q_{i,N-1} = 1/\sqrt{2}$ for $1 \leq i \leq N-1$ and $q_{i,j} = 0$ otherwise,
- $Q^{22} := diag_{inv}\left\{1, -i/\sqrt{2}, -i/\sqrt{2}, \cdots - i/\sqrt{2}\right\}$.

All in all, the Davies-Harte algorithm to simulate the increments of B^H can be summarised as follows.

Algorithm 25 The Davies-Harte algorithm (Wood and Chan, 1994).

i) Sample $Z \sim \mathcal{N}(0, I_{2N})$.

ii) Calculate $Z' = \Lambda_H^{1/2} Q Z$ where Q is defined in (7.52) with the $O(N)$ computational costs.

iii) Calculate $Z'' = PZ'$ with the $\mathcal{O}(N \log N)$ computational costs.

iv) The first N elements of $\delta^H Z''$ which gives G^H defined in (7.43).

Remark 22. *Perrin et al. (2002) prove that the circular matrix (7.47) is always non-negative definite for any $H \in (0,1)$, so $\{\lambda_k\}_{k=0}^{2N-1} \geq 0$ for any $H \in (0,1)$. Thus Algorithm 25 is well-defined for any $H \in (0,1)$. See also Craigmile (2003).*

Now we are ready to decouple dependency between B^H and H, indeed the Davies-Hartle algorithm (Algorithm 25) gives rise to the linear mapping which transforms $2N$ independent standard Gaussian variables into fGN at N discrete time instances. That is, we have that:

$$B^H = F(Z, H). \tag{7.53}$$

Therefore, instead of using directly X, we can treat and use Z as a latent variable, and now Z and H are a-priori independent. As a result, the full dependency between X and H is now decoupled. We summarise the dependencies among the involved variables in the models being studied as the hierarchical graph in Figure 23.

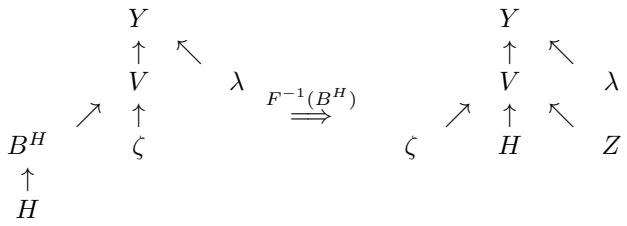


Figure 23: Dependency structures of the model. The left hierarchical graph shows the dependency structure of the model before we apply the transform $F^{-1}()$ in (7.53) and the right one shows the dependency structure of the model after we apply such a transform. The notation $A \to B$ should be understood as the variable B depends on the variable A.

We end this section by noting about an interpretation of the solution of (7.4) and corresponding numerical scheme. As shown by Sussmann (1978), for scalar B^H, one can obtain solutions for (7.4) in the sense that, given any fixed $t \mapsto B_t^H(\omega)$, one can define solutions for (7.4) any continuously differentiable paths in a a small neighbourhood of $B_t^H(\omega)$ for any $H \in (0,1)$, and then one can define the solution of (7.4) as the limit of such solutions as the neighbourhood gets tighter. This interpretation is often called the Doss-Sussmann interpretation in the literature. We note that when $H = \frac{1}{2}$, the Doss-Sussmann interpretation coincides with the classical Stratonovich interpretation.

As for numerical solutions corresponding the Doss-Sussmann interpretation, special care is needed.

Recall that we have defined $t_j = jT/N$ for $j = 0, 1, \cdots N$ with the simplified index $t_j = j$. Given simulated path of B^H by Algorithm 25, we apply the Euler-Maruyama scheme to discretize (7.4), for $j = 0, 1, \cdots N$:

$$V_j = V_{j-1} + b_\zeta(V_{j-1})\Delta t + \sigma_\zeta(V_{j-1})(B_j^H - B_{j-1}^H), \tag{7.54}$$

where $\Delta t := t_j - t_{j-1}$ and $V_0 = 0$. As shown in Lysy and Pillai (2013), such Euler-Maruyama scheme applied to B^H-driven multiplicative stochastic integrals will diverge to infinity when $H < 1/2$. Indeed, it can be shown that the quadratic variation of fBm explodes when $H < 1/2$ (Lysy and Pillai, 2013). Therefore, to discretize (7.4) by the Euler-Maruyama scheme, we may need to restrict our attention to a particular family of models, and remove $\sigma_\zeta(\cdot)$ from models, as suggested by Lysy and Pillai (2013). Critically, under the Doss-Sussmann interpretation, standard stochastic calculus rules can be applied for any $H \in (0, 1)$. That is, we can pick a sufficiently smooth mapping h and for the process $h(X_t)$, we can apply the standard change of variables rule under the Doss-Sussmann interpretation:

$$h(V_t) = V_0 + \int_0^t \nabla_x h(V_s) b_\zeta(V_s) ds + \int_0^t \nabla_x h(V_s)\sigma_\zeta(V_s) dB_s^H, \tag{7.55}$$

this is consistent with the change of variable formula for Stratonovich integrals. In particular, consider the following process:

$$A_t = h(V_t) := \int_{v_0}^{V_t} \frac{1}{\sigma_\zeta(u)} du, \tag{7.56}$$

this transformation is can be understood as a version of the Lamperti transformation. Then applying the standard Ito's lemma gives rise to:

$$dA_t = b_\zeta^A(A_t)dt + dB_t^H, \tag{7.57}$$

where we have defined $b_\zeta^V(A_t) := \frac{b_\zeta(h^{-1}(A_t))}{\sigma_\zeta(h^{-1}(A_t))}$. Equivalently, we can write (7.57) as $dA_t = \left(\frac{b_\zeta(V_t)}{\sigma_\zeta(V_t)}\right) dt + dB_t^H$. The standard Euler-Maruyama scheme for the transformed process in (7.57) will then converge to the analytical solution in an appropriate mode, under regularity conditions. Thus now we can apply the standard Euler-Maruyama scheme, and which is summarised as follows.

Algorithm 26 The standard Euler-Maruyama scheme for (7.4)

i) Simulate the increments of B^H by Algorithm 25.

ii) For $i = 1, \cdots N$, calculate:

$$A_i = A_{i-1} + b_\zeta^A(A_{i-1})\Delta t + (B_i^H - B_{i-1}^H),$$

with $A_0 = 0$, where $b_\zeta^A(A_t) := \frac{b_\zeta(h^{-1}(X_t))}{\sigma_\zeta(h^{-1}(X_t))}$ and the transformation $h(\cdot)$ is defined in (7.56).

iii) Return $V_i = h^{-1}(A_i)$ for $j = 0, 1, \cdots N$ with $V_0 = v_0$.

Remark 23. *Although the Doss-Sussmann interpretation can be used also for multi-dimensional SDEs, Algorithm 26 can in principal be followed for general models with a scalar scalar differential equation and B^H. For multi-dimensional models, one cannot avoid multiplicative stochastic integrals. For $H < 1/2$, one can use other interpretations for solving SDEs with corresponding numerical schemes such as a Milstein-type scheme, see Deya et al. (2012); Neuenkirch et al. (2010) for instance*

7.4.2 Validity of Advanced HMC for fBM models.

Let $\text{Leb}^{\otimes d}$ denote the $d-$dimensional Lebesgue measure. Then the joint posterior density of (Z, θ) of interest w.r.t. the reference measure $\otimes_{i=1}^{2N} N(0,1) \otimes \text{Leb}$ is given by:

$$\frac{d\Pi^N}{d\left(\otimes_{i=1}^{2N} N(0,1) \otimes \text{Leb}^{\otimes d}\right)}(Z, \theta \mid Y) \propto p^N(Y \mid Z, \theta)\pi_0(\theta). \tag{7.58}$$

That is, one can define the target of interest as a change of measure from a Gaussian law. Here N superscript is used to emphasise the fact that the joint posterior density in (7.58) is obtained as an $N-$dimensional proxy for the infinite-dimensional path V via Algorithm 26. Then, the target density can be now expressed as (again notice that Z is the $2N-$dimensional standard Gaussian vector):

$$\eta_n^N(Z, \theta \mid Y) \propto \exp\left(-\frac{1}{2}\langle Z, Z\rangle - \Phi_n(Z, \theta)\right), \tag{7.59}$$

where we have defined:

$$\Phi_n(Z, \theta) := -\log \pi_0(\theta) - \log p^N(Y \mid Z, \theta)^{\phi_n}, \tag{7.60}$$

so that Assumption 18 holds. This lead us to apply Algorithm 6 to to update jointly (Z, θ). We note that the advanced HMC sampler that jointly updates (Z, θ) cannot be derived directly from the advanced HMC algorithm. We set $x = (z, \theta)$, and now the corresponding Hamiltonian equations also can be split into:

$$\begin{cases} \frac{dx}{d\tau} = \frac{\partial \mathsf{H}_1}{\partial v} = 0, & \frac{dv}{d\tau} = -\frac{\partial \mathsf{H}_1}{\partial x} = -M^{-1}\nabla\Phi_n(x). \\ \frac{dx}{d\tau} = \frac{\partial \mathsf{H}_2}{\partial v} = v, & \frac{dv}{d\tau} = -\frac{\partial \mathsf{H}_2}{\partial x} = -(z, 0)^\top. \end{cases} \tag{7.61}$$

Therefore, we can also define the corresponding solution operators $\tilde{\Xi}_\tau^1$ and $\tilde{\Xi}_\tau^2$ analytically:

$$\tilde{\Xi}_\tau^1 := \left(x, v - \tau M^{-1}\nabla\Phi_n(x)\right), \tag{7.62}$$

$$\tilde{\Xi}_\tau^2 := \left((\cos(\tau)z + \sin(\tau)v_z, \theta + \tau v_\theta), (-\sin(\tau z + \cos(\tau)v_z, v_\theta)\right). \tag{7.63}$$

respectively. As a result, now we can carry out Algorithm 6 based on the L iterative steps in (7.41) as before.

To see its validity, recall that we have an $N-$dimensional proxy (7.59) of the true target posterior (3.1) defined on an infinite-dimensional space. Therefore, it is critically important to see that Algorithm 6 is a well-defined algorithm in the limit $N \to \infty$. To be precise, we want Algorithm 6 to leave

the target (3.1) invariant, and has mesh-free mixing with increasing N given fixed (ϵ, T). To check such properties following closely Beskos et al. (2015, 2013a), now we treat z as an infinite-dimensional *i.i.d.* standard Gaussian random vector, that is $z \in \mathbb{R}^\infty$, and we also have the parameters $\theta \in \mathbb{R}^d$. Then, the target distribution (3.1) corresponding to z, θ and Y can be assumed to be defined on the infinite-dimensional space $\mathcal{H} := \mathbb{R}^\infty \times \mathbb{R}^d$ via the following change of measure:

$$\frac{d\Pi_n}{d\left\{\otimes_{i=1}^\infty \mathcal{N}(0,1) \times \mathrm{Leb}^{\otimes d}\right\}}(Z, \theta \mid Y) \propto \exp\left(-\Phi_n(Z, \theta)\right), \tag{7.64}$$

for the function $\Phi : \mathcal{H} \mapsto \mathbb{R}$ defined in (7.60), where again Leb stands for the the Lebesgue measure. We also need the infnite-dimensional vector of partial derivatives (Fréchet sense) $\nabla\Phi : \mathcal{H} \mapsto \mathcal{H}$. Then, we have the corresponding velocity $v = (v_z, v_\theta) \in \mathcal{H}$. In addition, let $\mathcal{M}$ be the corresponding generalisation of the mass matrix M in Algorithm 5 whose upper-left block is now an infinite-dimensional identity matrix I_∞ instead of I_{2N}. That is, $\mathcal{M} : \mathcal{H} \mapsto \mathcal{H}$ is the linear operator $(z, \theta)^\top \mapsto \mathcal{M}(z, \theta)^\top = (z, A\theta)^\top$. Notice that we also have corresponding $\tilde{\Xi}_\tau^1$, $\tilde{\Xi}_\tau^2$, $\Psi_\epsilon : \mathcal{H} \times \mathcal{H} \mapsto \mathcal{H} \times \mathcal{H}$.

Consider the joint law on (x, v) denoted by $Q(dx, dv) := \Pi_n(dx) \otimes \mathcal{N}(0, \mathcal{M}^{-1})(dv)$. The key observation of the following proposition is that the leapfrog mapping Ψ_ϵ projects $(x_0, v_0) \sim Q(dx, dv)$ to (x_ϵ, v_ϵ) which has a distribution that is absolutely continuous w.r.t. $Q(dx, dv)$. This property implies the existence of a non-zero acceptance probability even when $N = \infty$ corresponding to the current infinite-dimensional set-up. This is clear for $\tilde{\Xi}_\epsilon^2$ since it just performs a rotation on (z, v_z) which is invariant for $\otimes_{i=1}^\infty \mathcal{N}(0,1) \times \otimes_{i=1}^\infty \mathcal{N}(0,1)$, see Neal (2011) for instance. Therefore, the mapping $\tilde{\Xi}_\epsilon^2$ preserves the absolute continuity property of $Q(dx, dv)$. To ensure that also the mapping $\tilde{\Xi}_{\epsilon/2}^1$ preserves the absolute continuity property of $v-$marginal of $Q(dx, dv)$, we need to assume that the the gradient $\nabla_z\Phi(z, \theta)$ should be in the Cameron-Martin space of $\otimes_{i=1}^\infty \mathcal{N}(0,1)$ for the translation $v \mapsto v - \frac{\epsilon}{2}\mathcal{M}^{-1}\nabla_z\Phi(x)$. This Cameron–Martin space is the one of squared summable infinite vectors, which we denote by ℓ_2, see Da Prato (2006, Chapter 1) for instance. Thus, we make the following assumption.

Assumption 19. *Let ℓ_2 be the Cameron-Martin space of $\otimes_{i=1}^\infty \mathcal{N}(0,1)$ which is the space of squared summable infinite vectors. Then $\nabla_z\Phi(z, \theta) \in \ell_2$ for all $\xi \in \mathcal{H}$ w.p.1 under $\otimes_{i=1}^\infty \mathcal{N}(0,1) \times p(d\theta)$.*

For a further study, we define the following reference measure on (x, v):

$$Q_0 = Q_0(dx, dv) := \left\{\otimes_{i=1}^\infty \mathcal{N}(0,1) \times \mathrm{Leb}^{\otimes d}\right\}(dx) \otimes \mathcal{N}(0, \mathcal{M}^{-1})(dv), \tag{7.65}$$

so that the joint target distribution is expressed as $Q(dx, dv) \propto \exp\left(-\Phi_n(x)\right) Q_0(dx, d\theta)$. Then we consider the sequence of probability measures $Q^{(k)} = Q \circ \Psi_\epsilon^{-\circ(k)}$ for $1 \le k \le L$. This $Q^{(k)}$ can be understood as the push-forward projection flow of the target measure $Q(dx, dv)$ via the leapfrog steps. For $\mathsf{H}(x, v)$, we define the difference as $\Delta\mathsf{H}(x_0, v_0) := \mathsf{H}(x_0, v_0) - \mathsf{H}(x_L, v_L)$ with the straightforward extension to $\mathbb{R}^\infty$of the inner product involved. Note that even if $\mathsf{H}(x_0, v_0) = \infty$, the difference $\Delta\mathsf{H}(x_0, v_0)$ does not explode (Beskos et al., 2015). Assuming stationarity, so that $(x_0, v_0) \sim Q$. Recall that $(x_L, v_L) = \Psi_\epsilon^{\circ(L)}(x_0, v_0)$. Then, we can write for the next position $x^\star$ of the Markov chain induced

by Algorithm 6:

$$x^\star = \mathbb{I}\left\{u \leq \alpha_{\mathsf{H}}\left(\Psi_\epsilon^{-\circ(L)}(x_T, v_T)\right)\right\} x_T + \mathbb{I}\left\{u > \alpha_{\mathsf{H}}\left(x_0, v_0\right)\right\} x_0, \tag{7.66}$$

where $u \sim Unif[0,1]$, $\mathbb{I}$ denotes the indicator function.

Proposition 43. *Assume that Assumption 19 holds. Then we have that:*

i) $Q^{(L)}$ *is is absolutely continuous w.r.t.*Q_0 *with the density* $\frac{dQ^{(L)}}{dQ_0}(x_L, v_L) = \exp\left(\Delta\mathsf{H}(x_0, v_0) - \Phi_n(x_L)\right)$.

ii) The Markov chain in (7.66) *has the invariant distribution* $\Pi_n(x)$ *defined in* (7.64).

Proof. Following Proposition 21, Beskos et al. (2015, Proposition 1) show that $Q^{(L)}$ is absolutely continuous w.r.t. Q_0 in (7.65) with the probability density $\frac{dQ^{(L)}}{dQ_0}(x_L, v_L) = \exp\left(\Delta\mathsf{H}(x_0, v_0) - \Phi_n(x_L)\right)$ by making use of the Cameron-Martin formula, see e.g Da Prato (2006, Theorem 2.8) and Lemma 7. Using this, it can be shown that, for any bounded and continuous test function $f : \mathcal{H} \mapsto \mathbb{R}$, $\mathbb{E}\left[f(\xi^\star)\right] = \mathbb{E}\left[f(\xi_0)\right]$ holds from Proposition 22. □

Remark 24. *Proposition 43 implies that the advanced HMC algorithm for joint inference* (Z, θ) *still possesses the mesh-free property, that is mixing properties do not deteriorate as* N *increases and* ϵ *remains fixed.*

7.4.3 Calculation of derivatives

Recall that Algorithm 6 involves the derivative $\nabla\Phi(Z, \theta)$. We note that Algorithm 25 and Algorithm 26 give rise to the composition $Z \mapsto G^H \mapsto A := \{A_1, \cdots A_N\}$. Assume that one can easily obtain the derivative $\nabla_\theta \log \pi_0(\theta)$. Then we have that:

$$\nabla_Z \log p^N(Y \mid Z, \theta)^{\phi_n} = \nabla_A \log p^N(Y \mid A, \theta)^{\phi_m} \left(\frac{dA}{dG^H}\right)^\top \left(\frac{dG^H}{dZ}\right)^\top, \tag{7.67}$$

$$\nabla_\zeta \log p^N(Y \mid Z, \theta)^{\phi_n} = \nabla_\zeta \log p^N(Y \mid A, \theta)^{\phi_n} + \nabla_A \log p^N(Y \mid A, \theta)^{\phi_n} \left(\frac{dA}{d\zeta}\right)^\top, \tag{7.68}$$

$$\nabla_\lambda \log p^N(Y \mid Z, \theta)^{\phi_n} = \nabla_\lambda \log p^N(Y \mid A, \theta)^{\phi_n}, \tag{7.69}$$

$$\partial_H \log p^N(Y \mid Z, \theta)^{\phi_n} = \nabla_A \log p^N(Y \mid A, \theta)^{\phi_n} \left(\frac{dA}{dG^H}\right)^\top \left(\frac{dG^H}{dH}\right)^\top, \tag{7.70}$$

where we have set:

$$\frac{dG^H}{dA} := \left\{\frac{\partial G_i^H}{\partial A_j}\right\}_{i,j} \in \mathbb{R}^{N\times N}, \quad \frac{dG^H}{dZ} := \left\{\frac{\partial G_i^H}{\partial Z_j}\right\}_{i,j} \in \mathbb{R}^{N\times 2N},$$
$$\frac{dA}{d\zeta} := \left\{\frac{\partial A_i}{\partial \zeta_j}\right\}_{i,j} \in \mathbb{R}^{N\times d}, \quad \frac{dG^H}{dH} := \left\{\frac{dG_i^H}{dH}\right\}_{i} \in \mathbb{R}^{N}.$$

Again recall that Algorithm 26 gives rise to $A_i = A_{i-1} + b_\zeta^A(A_{i-1})\Delta t + G_i^H$ for $i = 1, \cdots, N$. Then we set for $i = 2, \cdots, N$:

$$a_i = -1 - \nabla_A b_\zeta^A(A_{i-1})\Delta t,$$

as a consequence, we have that:

$$\frac{dG^H}{dA} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ a_2 & 1 & 0 & \cdots & 0 & 0 \\ 0 & a_3 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & a_N & 1 \end{bmatrix}. \tag{7.71}$$

Next, from Algorithm 25, we directly obtain:

$$\frac{dG^H}{dZ} = \mathfrak{proj}_{1:N,1:2N} \circ \delta^H P \Lambda^{1/2} Q, \tag{7.72}$$

where $\mathfrak{proj}_{1:N,1:2N}$ denotes the projection of the $2N \times 2N-$dimensional input matrix to its first N rows. Then we consider $\frac{dA}{d\zeta}$ which can be obtained recursively via starting from:

$$\nabla_\zeta A_1 = b_\zeta^A(A_0)\Delta t,$$

and for $i = 2, \cdots, N$:

$$\nabla_\zeta A_i = \nabla_\zeta A_{i-1} \times \left(1 + \nabla_A b_\zeta^A(A_{i-1})\Delta t\right) + b_\zeta^A(A_{i-1})\Delta t. \tag{7.73}$$

Following the Davies and Harte method (Algorithm 25), we have that:

$$\frac{dG^H}{dH} = \delta^H \mathfrak{proj}_{1:N} \circ P \frac{d\Lambda^{1/2}}{dH} QZ + \delta^H \log(\delta)\, \mathfrak{proj}_{1:N} \circ P\Lambda^{1/2} QZ. \tag{7.74}$$

From the analytic expression of $\{\lambda_k\}_{k=0}^{2N-1}$ the diagonal matrix $\Lambda^{1/2}$ in (7.48), we get:

$$\frac{d\lambda_k^{1/2}}{dH} = \frac{1}{2\lambda_k^{1/2}} \sum_{j=0}^{2N-1} \frac{dc_j}{dH} \exp\left(-2\pi i \frac{jk}{2N}\right),$$

where $\{c_j\}_{j=0}^{2N-1}$ are defined as the components of the first row of the circular matrix C in (7.47). Notice that $\frac{dc_j}{dH}$ can be easily obtained via the derivative of the lagged autocovariances of fBm w.r.t. H, $\frac{d\gamma(j)}{dH}$. where $\gamma(j)$ is defined as (7.44). From there, we have that:

$$\frac{d\gamma(j)}{dH} = \begin{cases} 0, & j = 0, \\ \log(2)2^{2H}, & j = 1, \\ (j+1)^{2H}\log(j+1) + (j-1)^{2H}\log(j-1) - 2\log(j)j^{2H}, & j \geq 2. \end{cases} \tag{7.75}$$

The expressions of $\nabla_\zeta \log p^N(Y \mid A, \theta)^{\phi_n}$, $\nabla_\lambda \log p^N(Y \mid A, \theta)^{\phi_n}$ and $\nabla_A \log p^N(Y \mid A, \theta)^{\phi_n}$ depend on specification of a model so we will derive them upon a model specification in the sequel.

7.5 Adaptive tuning strategies

The main advantage of using advanced HMC within SMC sampler is that one can calibrate adaptively the tuning parameters of the advanced HMC. Namely, after the resampling step in Algorithm 12, we have the equally weighted particles system $\left(\tilde{x}_{n-1}^{(i)}, \frac{1}{N}\right)_{i=1}^{N}$ which approximates $\hat{\eta}_n(dx_n)$ by construction. Recall that Algorithm 12 involves:

i) The mass matrix M of the proposal distribution $\mathcal{N}(0, M^{-1})$ for the velocity $v = (v_z, v_\theta) \in \mathbb{R}^{2N+d}$ in Algorithm 6.

ii) The sequence of temperatures $\{\phi\}_{n=0}^{p}$ for the the tempered posterior in (7.20).

iii) The number of steps L and the step size ϵ for the leapfrog integrator Ψ_τ in Algorithm 6.

7.5.1 The mass matrix M

Recall that for $x = (z, \theta)$, the mass matrix M has the form of:

$$M = \begin{pmatrix} I_{2N} & 0 \\ 0 & B \end{pmatrix},$$

here the identity sub-matrix I_{2N} is for v_z, and the sub-matrix $B = diag(b_1, \cdots b_d)$ is for v_θ. Since we have the resampled particles $\left(\tilde{x}_{n-1}^{(i)}, \frac{1}{N}\right)_{i=1}^{N}$ before we mutate $\tilde{x}_{n-1}^{(i)}$ via the advanced HMC kernel $\mathcal{K}(\tilde{x}_{n-1}^{(i)}, dx^{(i)})$, we can estimate the posterior mean and the variance of $\hat{\eta}_n(dx_n)$ respectively as follows:

$$\hat{\mu}_\theta(n) := \sum_{i=1}^{N} W_n^{(i)} \tilde{x}_{n-1}^{(i)}, \; \hat{\Sigma}_\theta(n) := \sum_{i=1}^{N} W_n^{(i)} \left(\tilde{x}_{n-1}^{(i)} - \hat{\mu}_\theta(n)\right)\left(\tilde{x}_{n-1}^{(i)} - \hat{\mu}_\theta(n)\right)^\top. \tag{7.76}$$

Namely, $\left(\tilde{x}_{n-1}^{(i)}, \frac{1}{N}\right)_{i=1}^{N}$ can be used to learn features of the target. Therefore, we use the following adaptive mass matrix at iteration n :

$$M_n := \begin{pmatrix} I_{2N} & 0 \\ 0 & diag\left(\hat{\Sigma}_\theta(n)\right) \end{pmatrix}. \tag{7.77}$$

This aprroach has been used also by Chopin (2002); Chopin et al. (2013); Jasra et al. (2011) in the context of an independent Metropolis-Hastings, particle markov chain monte carlo methods and inference for Lévy-driven stochastic volatility models for instance.

7.5.2 The sequence of temperature $\{\phi_n\}_{n=0}^{p}$

We follow the ideas in Beskos et al. (2016); Del Moral et al. (2012); Jasra et al. (2011). Given $\beta \in (0, 1)$, consider the following recursive equation:

$$\phi_n = \inf\left\{\phi_{n-1} < \phi \leq 1 : ESS(\phi_{n-1}) - \beta N = 0\right\}, \tag{7.78}$$

with $\phi_0 = 0$, where:

$$ESS(\phi_n) := \frac{1}{\sum_{i=1}^{N} \left(W_n^{(i)}\right)^2} \tag{7.79}$$

is the effective sample size (ESS) and $\alpha \in (0,1)$ is user specified value. The ESS has been commonly used to monitor the performance of SMC in the literature (Kong et al., 1994; Doucet and Johansen, 2009), and a common choice of β will be 0.5. Notice that the ESS takes a value between 1 and N since $\{W_n^{(i)}\}_{i=1}^N$ are normalized. For completeness, we use the convention that $\inf \emptyset = 1$. Under mild conditions, Beskos et al. (2016) show that a particle approximation of $\hat{\eta}_n(dx_n)$ with the sequence of temparatures solving (7.78) is still consistent, and the sequence $\{\phi_n\}_{n=0}^p$ solving (7.78) monotonically converges to 1. Notice that (7.78) can be easily computed by the bisection method, see e.g. Jasra et al. (2011); Beskos et al. (2016).

Algorithm 27 Adaptive learning of $\{\phi_n\}_{n=0}^p$

i) At the artificial time step $n-1$, calculate the ESS in (7.79).

ii) Choose the next inverse temperature ϕ_n by solving (7.78) via the bisection method.

7.5.3 The leapfrog integrator parameters (ϵ, L)

In practice, the performance of (advanced) HMC sensitively depends on a choice of (ϵ, L). Too large step size ϵ will yield low acceptance rates, as studied in Beskos et al. (2011). In contrast, too small step size will increase the computational cost. As for the number of the leap-frog L, too large L will give rise to the Hamiltonian trajectory which turns back towards its starting point, thus, wasting computational computational resource again. Also, too small L will induce random walk behaviour and the induced Markov chain might not be ergodic, as claimed in Neal (2011). To overcome these problems, Neal (2011) suggests that (ϵ, L) to be randomly sampled.

To do so, consider the following the expected Mahalanobis square jumping distance:

$$\mathsf{M}(x_{n-1}, x_n) := (x_{n-1} - x_n)^\top M_n^{-1} (x_{n-1} - x_n), \tag{7.80}$$

where M_n is defined in (7.77). This distance can be understood as the lag-1 integrated autocorrelation time and has been commonly used in the adaptive MCMC literature, see e.g. Andrieu and Thoms (2008); Sherlock and Roberts (2009). Let $h := (\epsilon, L)$. Then using (7.80), Fearnhead and Taylor (2013) propose that parameter h is to be chosen to maximise the following criterion:

$$g_n(h) := \int \hat{\eta}_{n-1}(d\tilde{x}_{n-1}) \mathcal{K}_n(\tilde{x}_{n-1}, dx_n) \mathsf{M}(\tilde{x}_{n-1}, x_n), \tag{7.81}$$

which is equivalent to minimising the average of the lag-1 integrated autocorrelation time. Notice that since after the selection step in Algorithm 12 one has the particle system $\left(\tilde{x}_{n-1}^{(i)}, \frac{1}{N}\right)_{i=1}^N$, h can be assigned to different particles at every time step n, and thus we can write $\left(h_n^{(i)}\right)_{i=1}^N$. The main idea of

Fearnhead and Taylor (2013) is that one can sample $\left(h_n^{(i)}\right)_{i=1}^N$ from some distribution, say $R_n(h_{n-1}^{(i)})$ which is independent from particles, and they are also weighted according to (7.80). As in Fearnhead and Taylor (2013), we use the following estimator of (7.81):

$$\widehat{\mathsf{M}}(x_{n-1}^{(i)}, x_n^{(i)}) := \frac{1}{\sqrt{L_{n-1}^{(i)}}} \mathsf{M}(\tilde{x}_{n-1}^{(i)}, \hat{x}_n^{(i)}) \times \alpha_{\mathsf{H}}^{(i)}, \tag{7.82}$$

where $\left\{\hat{x}_n^{(i)}\right\}_{i=1}^{N_\xi}$ are proposed particles based on the leapfrog integrator Ψ_ϵ and $\alpha_{\mathsf{H}}^{(i)}$ is defined in (7.41). We note that since $\left(\tilde{x}_{n-1}^{(i)}, \frac{1}{N}\right)_{i=1}^N$ approximates $\hat{\eta}_n(dx_n)$ by construction and $\left\{\hat{x}_n^{(i)}\right\}_{i=1}^N$ are accepted w.p. $\alpha_{\mathsf{H}}^{(i)}$, (7.82) can be understood as an unbiased estimate of (7.81). Then, as suggested by Fearnhead and Taylor (2013); Buchholz et al. (2018), we sample $\left(h_n^{(i)}\right)_{i=1}^{N_\xi}$ from:

$$\pi_n(h) \propto \sum_{i=1}^{N_\xi} \widehat{\mathsf{M}}(x_{n-1}^{(i)}, x_n^{(i)}) R_n(h_{n-1}^{(i)}), \tag{7.83}$$

$$R_n(h_{n-1}^{(i)}) := \mathcal{TN}(\epsilon; , \epsilon_{n-1}^{(i)}, 0.015^2) \otimes \left(\frac{1}{3}\mathbb{I}\left\{L_{n-1}^{(i)} - 1\right\} + \frac{1}{3}\mathbb{I}\left\{L_{n-1}^{(i)}\right\} + \frac{1}{3}\mathbb{I}\left\{L_{n-1}^{(i)} + 1\right\}\right), \tag{7.84}$$

where $\mathcal{TN}(a; , b, c)$ denotes a Gaussian distribution evaluated at a with mean b and the variance c truncated to positive real $\mathbb{R}_+$. That is, $\left(h_n^{(i)}\right)_{i=1}^{N_\xi}$ are sampled by first resampling $\left(h_{n-1}^{(i)}\right)_{i=1}^{N_\xi}$ w.p. (7.82) and next drawing from (7.84). Under some technical conditions, Theorem 4.1 of Fearnhead and Taylor (2013) proves that (7.83) converge in distribution to the Dirac mass measure centred on the point which attains the maximiser of (7.81) as $n \to \infty$. Therefore, sampling $\left(h_n^{(i)}\right)_{i=1}^N$ from (7.83) might guarantee sequential improvement of MCMC mixing with respect to n.

Algorithm 28 Adaptive sampling (ϵ, L)

i) After the mutation step in Algorithm 12 at the artificial time step $n-1$, calculate (7.82) for $i = 1, \cdots, N$.

ii) Resample $\epsilon_{n-1}^{(i)}$ and $L_{n-1}^{(i)}$ w.p. (7.82) for $i = 1, \cdots, N$.

iii) Sample $\epsilon_n^{(i)}$ and $L_n^{(i)}$ from (7.84) for $i = 1, \cdots, N$.

Algorithm 29 Adaptive SMC sampler with the advanced HMC.

i) At time $n = 0$, set $\phi_0 = 0$ and draw $z_0 \sim \otimes_{i=1}^{2N}\mathcal{N}(0,1)$ and $\theta_0 \sim \pi_0(\theta)$.

ii) Calculate the inverse temperature ϕ_n via Algorithm 27.

iii) Do from the step 1 to the step 4 of Algorithm 12.

iv) Calculate mass matrix M_n in (7.77).

v) Do the step 5 of Algorithm 12 based on M_n.

vi) Obtain $(\epsilon_{n+1}, L_{n+1})$ via Algorithm 28.

vii) Set $n \to n+1$ and repeat from the step 2 to the step 7 until ϕ_n=1.

7.6 Conclusion and remarks

Our main contribution of this section is that we have developed the novel MCMC based algorithm for a Bayesian model selection in the context of SDE models driven by fractional noise. Critically, compared with literature, our method can avoid successfully problems such as constructing a non-trivial transformation and the slow mixing by making use of pseudo marginal MCMC and HMC on a Hilbert space within SMC samplers. Besides, we have shown that how one can select adaptively tuning parameters using outputs from SMC samplers.

At the time of writing the thesis this section had not been completed. The first missing step we need to fulfil is to explore a way to correct the bias arising from adaptive sampling since an unbiased estimate of the evidence is the key to approximate exactly the posterior model probability. This could be done by running the algorithm twice. Namely, one could run the adaptive algorithm with arbitrary tuning parameters and then save all adaptively estimated. Then one could rerun the static algorithm with estimated tuning parameters. It could be shown that the difference between the first run and the second run indeed gives rise to an unbiased estimate. Beside, some numerical studies are also needed. For instance, comparing SV model with/without Hurst parameter would be interesting. Also, model selection for fractional SV and fractional GARCH would be interesting as well.

8 Summary and future directions

8.1 Summary of research

8.1.1 Asymptotic Analysis of Model Selection Criteria for General Hidden Markov Models

i) The paper obtains analytical results for the asymptotic properties of Model Selection Criteria for a general family of state space models (SSMs).

ii) We first derive AIC and BIC rigorously for SSMs. To the best of our knowledge, this is the first time to obtain such results.

iii) Next, we show that BIC is still strongly consistent for nested SSMs under conventional assumptions. Also, we show that AIC cannot be consistent for such models.

iv) Finally, we check our theoretical results through the empirical study. We confirm that our empirical study is consistent with the implication of theoretical results.

8.1.2 Online Smoothing for Diffusion Processes Observed with Noise

i) We introduce a novel particle algorithm for online estimation of smoothing expectations for a class of additive functionals, in the context of a rich family of diffusion processes with jumps.

ii) Our methodology is based on online particle smoothing on diffusion path space, and can be applied under the mild condition compared with the literature.

iii) We overcome the unavailability of the transition density of the underlying SDE by working on the augmented path space.

iv) Finally, we apply our methods to online parameter inference and model selection for the real data and the problems motivated by the finance literature.

8.1.3 Adaptive Bayesian Model Selection for Diffusion Models

i) We introduce a general framework of Bayesian model selection based on Sequential Monte Carlo sampler and Hamiltonian Monte Carlo on high dimensional spaces.

ii) The proposed algorithm is well-defined on a functional space so that the speed of mixing does not depend on the dimension of the models being considered.

iii) Our methodology can be considered as an exact approximation of the posterior model probability therefore is has several advantages. For instance we can avoid potential problems arising from trans-dimensional MCMC methods.

iv) At the time of writing the thesis section this had not been completed. We will apply our methods to fractional stochastic volatility models has become popular to model derivative securities, such as options.

8.2 Future directions

8.2.1 Robust adaptive sequential Monte Carlo methods for non-Markovian state space models

There has recently been an interesting line of research when it comes to merging deep learning architectures with stochastic models (Louizos et al., 2017; Rangapuram et al., 2018). Importantly, some of these models can be understood as (Gaussian) non-Markovian state space models. Sequential Monte Carlo (SMC) is a particularly useful sequential (and online in many cases) method to approximate intractable functions and estimate parameters of SSMs, but the Markovian structure of a target model is often critical. To make SMC applicable for non-Markovian state space models, one needs to make a proposal or a target itself taking into account global information of the target distribution. In particular, I am interested in the latter approach based on the robust divergences, for instance, $\beta-$divergence (Basu et al., 1998). Considering β is a sequence such that $\beta_0 = \infty > \beta_1 \cdots > \beta_n = 0$ gives rise to the tempered sequence of distributions, say $\{\eta_t(d\theta)\}_{t=0}^n$, on a common measurable space. SMC samplers (Del Moral et al., 2006) allows offline Bayesian inference to be conducted for such distributions. Upon developing an adaptive way to find a reasonable value of $\{\beta_t\}_{t=0}^n$, this approach will help robust adaptive approximate Bayesian inference for such non-Markovian models.

8.2.2 Deep learning with general Bayesian principles

Applying Bayesian inference to deep learning has been recently becoming a centre of interest in a machine learning community (Gal, 2016) to quantify uncertainty. Bayesian inference is an optimal updating procedure as long as a statistical model is correctly specified, or no connection between any data and parameters. Interestingly, Bissiri et al. (2016) argue that the posterior minimising the expected loss criteria $\pi(x \mid \theta) \propto \exp(-\omega \ell_\theta(x))\pi(d\theta)$ is still valid and optimal updating procedure even if model is misspecified, where ω is weight, $\ell_\theta(x)$ is a loss function (not necessary log likelihood) and $\pi(d\theta)$ is a prior distribution. It is important that the weight ω is calibrated carefully, and Holmes and Walker (2017) study how to select ω when the model is the exponential family. It turns out that merging such a general Bayesian inference with VOGN (Khan et al., 2018), for instance, will give rise to an online variational method for deep learning with robustness over model misspecification. Also, using a robust divergence in a loss function $\ell_\theta(x)$ will make deep learning stable against outliers.

8.2.3 Speeding up MCMC with intractable likelihood functions

One of the main difficulties of MCMC comes from its computational cost. For instance, implementing the Metropolis-Hastings algorithm involves evaluation of the likelihood ratio, which is too costly an operation, especially for big data so that speeding up MCMC is a critical problem. Several techniques are available towards accelerating MCMC based on subsampling and control variates (Bardenet et al., 2017; Quiroz et al., 2019) to reduce the cost. However, they cannot be directly applied to models with intractable likelihood functions since the methods require the analytical expression of them. Merging approximate Bayesian inference with MCMC is called noisy MCMC (Alquier et al., 2016), and this will be an effective way to sample from such models. In particular, inspired by the literature, I am interested in establishing concentration inequalities between noisy likelihood functions and exact ones.

Also, developing an algorithm to construct an efficient proxy of the noisy likelihood ratio might be needed to control the variance of the ratio. Such inequalities and proxy will provide ways to speed up noisy MCMC.

8.2.4 The No-U-Turn sampler on functional spaces

Hamiltonian Monte Carlo (HMC) is a particular class of Markov chain Monte Carlo methods, which is based on the Hamiltonian dynamics to sample from a target distribution. In practice, HMC involves a numerical integrator to approximate the Hamiltonian dynamics so that some tuning parameters are also required. Hoffman and Gelman (2014) introduce the No-U-Turn sampler which adaptively selects the tuning parameters of HMC, and has become a de facto standard of HMC. Critically, many interesting models can be expressed as a change of Gaussian measure on a functional space. Therefore, developing a well-defined No-U-Turn sampler on functional spaces will help us to sample efficiently from a target defined on a high dimensional space such as diffusion models, Bayesian non-parametric models and deep neural networks. This is because the mixing speed of such well-defined HMC will not depend on the dimension of targets (Beskos et al., 2011; Cotter et al., 2013).

8.2.5 Asymptotic analysis of the Monte Carlo MLE for SSMs with SMC outputs

In general, the likelihood function of SSMs is not analytically available so that it should be approximated by, for instance, SMC. Therefore, what we can obtain is the Monte Carlo MLE (Geyer, 1994). The main difficulty comes from the fact that SMC outputs are not continuous w.r.t. inputs due to resampling so that one can only obtain point-wise convergence results. To obtain strong consistency and asymptotic normality of the MLE, one needs to show that some uniform convergence results. It turns out, although asymptotic analytical results are available (Douc et al., 2004), these results cannot be directly applied to the problem. Therefore, establishing such results is quite important and challenging. To do so, one first needs to develop SMC for SSMs with a continuous likelihood, for instance, based on Malik and Pitt (2011). Then one needs to show that the Monte Carlo MLE converges to the true MLE and identify the optimal choice of SMC iterations. Compared with the literature, e.g. Beskos et al. (2009), such analysis will be harder since the resampling step of SMC creates dependency. Strong law of large numbers in a separable Banach space might be useful, see Appendix C and Beskos et al. (2009).

A Taylor's theorem with the exact integral form of the remainder

Taylor's theorem may be one of the most useful tools for not only analysis but also statistics. For instance, in the context of statistics, one might often want to expand the log-likelihood evaluated at an estimate $\hat{\theta}_n$ around the true parameter, say $\theta^\star$. Let X be a open subset of $\mathbb{R}^p$ with $p \geq 1$. Also let $f : X \to \mathbb{R}^q$ with $q \geq 1$ be continuously differentiable with $p \times q$ derivative matrix ∇f. As Feng et al. (2014) pointed out, for $a, b \in X$, *there does not in general exist* a θ on the line segment between a and b such that:

$$f(b) = f(a) + \nabla f(\theta)(b - a).$$

Then the following Taylor's theorem with the exact integral form of the remainder (Lang, 2012, Section14) may be useful.

Theorem 22. *Let U be open in E and $f : U \to F$ be of class C^p, that is first p derivatives all exist and are continuous. Let $x \in U$ and $y \in E$ such that the segment $x + ty$, $0 \leq t \leq 1$, is contained in U. Denote $y^{(k)}$ by the k-tuple $(y, y, \cdots, y)$. Then we have that:*

$$f(x + y) = f(x) + Df(x)y + \cdots + \frac{D^{p-1} f(x) y^{(p-1)}}{(p-1)!} R_n,$$
$$R_n = \int_0^1 \frac{(1-t)^{p-1}}{(p-1)!} D^p f(x + ty) y^{(p)} dt,$$

where $Df : U \to \mathcal{L}(E, F)$ is the differential operator acts on the set of linear mappings defined over U and taking values in F, see Lang (2012, Section 14) for details.

As an example, consider the MLE $\hat{\theta}_n$ and the the score function $\nabla \ell_n(\theta)$. Again, let $\theta^\star$ be the true parameter. Take $x = \hat{\theta}_n$ and $y = \hat{\theta}_n - \theta^\star$ in Theorem 22. Then, the first order expansion of $\nabla \ell_n(\hat{\theta}_n)$ around $\theta^\star$ with the exact integral form of the remainder is given by:

$$0 = \nabla \ell_n(\hat{\theta}_n) = \nabla \ell_n(\theta^\star) + (\hat{\theta}_n - \theta^\star)^\top \int_0^1 \nabla \ell_n(\theta^\star + t(\hat{\theta}_n - \theta^\star)) dt.$$

B Some asymptotic results for the class of estimators

Let $\{y_n\}_{n \geq 0}$ be a strongly stationary and ergodic process which is parametrized by $\theta \in \Theta$. In many cases, the parameter θ is estimated as the solution to:

$$\hat{\theta}_n^E := \arg\max_{\theta \in \Theta} Q_n(\theta). \tag{B.1}$$

This class of estimators is often called *extreme estimator*, see e.g., Van der Vaart (2000). For instance, the MLE, the M-estimator and the method of moments estimator are extreme estimators. Let $\theta^\star$ be the true parameter. The the following conditions are known as the sufficient conditions for $\hat{\theta}_n^E \to \theta^\star$ in probability.

Theorem 23. *Let $\mathcal{B}_\delta(\theta^\star) := \{\theta \in \Theta : |\theta^\star - \theta| < \delta\}$. Assume that there exist a non-stochastic function $Q(\theta)$ and $\theta^\star \in \Theta$ such that:*

i) For any $\delta > 0$, $\sup_{\theta \in \Theta \setminus \mathcal{B}_\delta(\theta^\star)} Q(\theta) < Q(\theta^\star)$.

ii) $\lim\sup_{n\to\infty} \sup_{\theta\in\Theta} |Q_n(\theta) - Q(\theta)| = 0$ in probability.

Then $\hat{\theta}_n^E \to \theta^*$ in probability.

Proof. Let $\hat{\theta}_n^E \in \Theta \setminus \mathcal{B}_\delta(\theta^\star)$. Then there exists $\epsilon > 0$ such that $Q(\theta^\star) - Q(\hat{\theta}_n^E) \geq \epsilon$. Thus we have that for any $\delta > 0$, $\mathbb{P}\left(\hat{\theta}_n^E \in \Theta \setminus \mathcal{B}_\delta(\theta^\star)\right) \leq \mathbb{P}\left(Q(\theta^\star) - Q(\hat{\theta}_n^E) \geq \epsilon\right)$, and:

$$\begin{aligned}
\mathbb{P}\left(Q(\theta^\star) - Q(\hat{\theta}_n^E) > \epsilon\right) &= \mathbb{P}\left(Q(\theta^\star) - Q_n(\theta^\star) + Q_n(\theta^\star) - Q_n(\hat{\theta}_n^E) + Q_n(\hat{\theta}_n^E) - Q(\hat{\theta}_n^E) \geq \epsilon\right) \\
&\leq \mathbb{P}\left(\sup_{\theta\in\Theta} |Q(\theta) - Q_n(\theta)| + 0 + \sup_{\theta\in\Theta} |Q(\theta) - Q_n(\theta)| \geq \epsilon\right) \\
&= \mathbb{P}\left(2\sup_{\theta\in\Theta} |Q(\theta) - Q_n(\theta)| \geq \epsilon\right).
\end{aligned}$$

Then by the assumption, we have that $\lim_{n\to\infty} \mathbb{P}\left(2\sup_{\theta\in\Theta} |Q(\theta) - Q_n(\theta)| \geq \epsilon\right) = 0$. □

Lemma 16. *Under the following assumptions:*

i) $Q(\theta)$ is uniquely maximized at $\theta^\star$.

ii) $Q(\theta)$ is continuous on Θ.

iii) Θ is a compact set.

Then for any $\delta > 0$, we have that $\sup_{\theta \in \Theta \setminus \mathcal{B}_\delta(\theta^\star)} Q(\theta) < Q(\theta^\star)$.

Proof. Assume that $\sup_{\theta \in \Theta \setminus \mathcal{B}_\delta(\theta^\star)} Q(\theta) = Q(\theta^\star)$. The there exists a sequence $\{\theta_n\}$ such that $\theta_n \in \Theta \setminus \mathcal{B}_\delta(\theta^\star)$ for any n and $\lim_{n\to\infty} Q(\theta_n) = Q(\theta^\star)$. Since Θ is a compact set, we can always find a subsequence $\{\theta_{n_i}\}$ converging to a limit $\theta^* \in \Theta \setminus \mathcal{B}_\delta(\theta^\star)$ as $i \to \infty$. Since $Q(\theta)$ is continuous on Θ, we have that $\lim_{i\to\infty} Q(\theta_{n_i}) = Q(\theta^*)$, and this implies that $Q(\theta^\star) = Q(\theta^*)$. This is a contradiction since $\theta^* \in \Theta \setminus \mathcal{B}_\delta(\theta^\star)$ and $Q(\theta)$ is uniquely maximised at $\theta^\star$ by the assumption. □

Consider the M-estimator. In this case, $Q_n(\theta)$ is often of the form $\frac{1}{n}\sum_{i=1}^n m(y_i, \theta)$ and $Q(\theta) = \mathbb{E}[m(y_i, \theta)]$ with some known function $m(y_i, \theta)$. Then the following might be useful.

Lemma 17. *Suppose that, w.p.1, $\theta \mapsto m(y_i, \theta)$ is continuous for any y_i and $\mathbb{E}\left[\sup_{\theta\in\Theta} |m(y_i, \theta)|\right] < \infty$. Then $\mathbb{E}[m(y_i, \theta)]$ is continuous on Θ w.p.1.*

Proof. By the assumptions, we have that $\lim_{\theta'\to\theta} m(y_i, \theta')$ w.p.1. and $|m(y_i, \theta)| \leq \sup_{\theta\in\Theta} |m(y_i, \theta)|$, the results follow from the dominated convergence theorem. □

C Strong law of large numbers in a separable Banach space

Let θ_n^E be an extreme estimator. To prove $\theta_n^E \to \theta^\star$ w.p.1, the following will be needed:

$$w.p.1.\ \lim_{n\to\infty} \sup_{\theta\in\Theta} |Q_n(\theta) - Q(\theta)| = 0. \tag{C.1}$$

For instance, let X_i be *i.i.d.* random variables, and take $Q_n(\theta) = \frac{1}{N}\sum_{i=1}^N X_i(\theta)$ such that $\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^N X_i(\theta) - Q(\theta)\right] = 0$ for any $\theta \in \Theta$. Then Alamogordo's Strong Law of Large Numbers implies that for any $\theta \in \Theta$, $\frac{1}{N}\sum_{i=1}^N X_i(\theta) \to Q(\theta)$ w.p.1 holds. Also assume that the MLE $\hat{\theta}_n \to \theta^\star$ w.p.1. Then we have that:

$$\begin{aligned}
&\left|\frac{1}{N}\sum_{i=1}^N X_i(\theta) - Q(\theta^\star)\right| \\
\leq &\left|\frac{1}{N}\sum_{i=1}^N X_i(\theta) - Q(\hat{\theta}_n)\right| + \left|Q(\hat{\theta}_n) - Q(\theta^\star)\right| \\
\leq &\sup_{\theta\in\Theta}\left|\frac{1}{N}\sum_{i=1}^N X_i(\theta) - Q(\hat{\theta}_n)\right| + \left|Q(\hat{\theta}_n) - Q(\theta^\star)\right|.
\end{aligned}$$

If $\theta \mapsto Q(\theta)$ is continuous, then $\lim_{n\to\infty}\sup_{\theta\in\Theta}\left|\frac{1}{N}\sum_{i=1}^N X_i(\theta) - Q(\hat{\theta}_n)\right| = 0$ w.p.1 implies $\frac{1}{N}\sum_{i=1}^N X_i(\hat{\theta}_n) \to Q(\theta^\star)$ w.p.1.

To prove (C.1), strong law of large numbers on a separable Banach space might be useful. Let $(\mathbf{B}, \|\|)$ be a real separable Banach space and $\mathfrak{X}_i$ are independent $\mathbf{B}$-valued random variables . Then we have the following.

Theorem 24. *Azlarov and Volodin (1982); Mourier (1953); Hoffmann-Jorgensen and Pisier (1976). Let $(\mathbf{B}, \|\|)$ be a real separable Banach space and $\mathfrak{X}_i$ are independent $\mathbf{B}$-valued random variables . Then we have that:*

$$w.p.1.\ \lim_{N\to\infty}\left\|\frac{1}{N}\sum_{i=1}^N \mathfrak{X}_i\right\| = 0,$$

under the following assumptions:

i) $\mathbb{E}\left[\|\mathfrak{X}_i\|\right] < \infty$.

ii) $\mathbb{E}\left[\mathfrak{X}_i\right] = 0$.

Proposition 44. *Consider $Q_n(\theta)$ is constructed by some independent random variables X_n. Assume that:*

i) The parameter space Θ is compact.

ii) $\theta \mapsto Q(\theta)$ is continuous.

iii) For any $\theta \in \Theta$, $\mathbb{E}\left[Q_i(\theta, X_i)\right] = Q(\theta)$.

iv) W.p.1, $\theta \mapsto Q_i(\theta, X_i)$ is continuous.

v) W.p.1, $\sup_{\theta\in\Theta} \mid Q_i(\theta, X_i) \mid < \infty$.

*Then w.p.*1. $\lim\sup_{n\to\infty}\sup_{\theta\in\Theta} |Q_n(\theta) - Q(\theta)| = 0$ *holds.*

Proof. Take $\mathbf{B}$ be the space of continuous real functions on the compact Θ and $\|\|$ be the uniform norm, that is $\|f\| = \sup_{\theta\in\Theta} \mid f(\theta) \mid$. Then such $(\mathbf{B}, \|\|)$ is a real separable Banach space. By the assumptions, $Q(\theta) \in \mathbf{B}$ and $Q_n(\cdot, X_i)$ takes values in $\mathbf{B}$, also $\mathbb{E}\left[\|Q_n(\theta, X_i)\|\right] < \infty$ and $\mathbb{E}\left[Q_n(\cdot, X_i) - Q(\theta)\right] = 0$. Then applying Theorem 24 to $Q_n(\cdot, X_i) - Q(\theta)$ gives rise to the claim. □

Theorem 24 might hold for a martingale difference sequence (Hoffmann-Jorgensen and Pisier, 1976) and a stationary (not necessarily ergodic) stochastic process (Cuny, 2015; Dedecker and Merlevède, 2008).

D Convergence of moments and uniform integrability

Suppose that a random variable X_n converges to X in distribution, e.g. $X_n \Rightarrow X$. In general, convergence in distribution does not imply convergence in moments, that is $X_n \Rightarrow X$ does not mean $\mathbb{E}\left[X_n\right] \to \mathbb{E}\left[X\right]$. However, if X_n are uniformly integrable, then convergence in distribution implies convergence in moments. We refer to Billingsley (2008) for instance.

Definition 32. *The sequence* X_n *is said to be uniformly integrable if for any* $\epsilon > 0$ *there exists a constant* $C < \infty$ *such that* $\mathbb{E}\left[\mid X_n \mid \mathbb{I}\left\{\mid X_n \mid > C\right\}\right] < \epsilon$ *for any* n. *Or, equivalently,* $\lim_{C\to\infty} \sup_n \mathbb{E}\left[\mid X_n \mid \mathbb{I}\left\{\mid X_n \mid > C\right\}\right] = 0$.

Proposition 45. *If* $X_n \Rightarrow X$ *and* X_n *are uniformly integrable. Then* X *is integrable and* $\mathbb{E}\left[X_n\right] \to \mathbb{E}\left[X\right]$.

Proof. Since $X_n \Rightarrow X$, we can find random variables Y_n and Y defined on $((0,1), \mathcal{B}((0,1)), \text{Leb}((0,1)))$ such that Y_n has the same distribution as X_n, Y has the same distribution as X and $Y_n \to Y$ w.p.1. due to Skorokhod's representation theorem. Then applying Fatou's lemma to Y_n gives rise to $\mathbb{E}\left[\mid Y \mid\right] = \mathbb{E}\left[\liminf_n \mid Y_n \mid\right] \leq \liminf_n \mathbb{E}\left[\mid Y_n \mid\right]$. Since $\mid Y_n \mid$ and $\mid X_n \mid$ have the same distribution, we have that $\mathbb{E}\left[\mid Y_n \mid\right] = \mathbb{E}\left[\mid X_n \mid\right]$. Since X_n is uniformly integrable, we have:

$$\begin{aligned}\mathbb{E}\left[\mid X_n \mid\right] &= \mathbb{E}\left[\mid X_n \mid \mathbb{I}\left\{\mid X_n \mid > C\right\}\right] + \mathbb{E}\left[\mid X_n \mid \mathbb{I}\left\{\mid X_n \mid \leq C\right\}\right] \\ &\leq 1 + C.\end{aligned}$$

This implies that X is integrable since $\mathbb{E}\left[\mid Y \mid\right] = \mathbb{E}\left[\mid X \mid\right] \leq \mathbb{E}\left[\mid X_n \mid\right]$. The rest of the claim follows from theorem 25.12 of Billingsley (2008). □

The next claim known as the sufficient condition of Proposition 45.

Lemma 18. *Suppose that there exists* $\delta > 0$ *such that* $\sup_n \mathbb{E}\left[\mid X_n \mid^{1+\delta}\right] < \infty$. *Then* X_n *are uniformly integrable.*

Proof. We have that:

$$\sup_n \mathbb{E}\left[\mid X_n \mid \mathbb{I}\{\mid X_n \mid > C\}\right] \leq \sup_n \mathbb{E}\left[\frac{\mid X_n \mid^{1+\delta}}{C^\delta}\mathbb{I}\{\mid X_n \mid > C\}\right]$$
$$\leq \sup_n \mathbb{E}\left[\frac{\mid X_n \mid^{1+\delta}}{C^\delta}\right] \to 0, \ as \ C \to \infty.$$

□

E L^2-bound for Monte Carlo estimates

Proposition 46. *Assume the* $x^{(i)} \overset{i.i.d.}{\sim} \mu \in \mathcal{P}(E)$. *Then, for any* $f \in \mathcal{B}_b(E)$, *we have that:*

$$\sup_{\|f\|_\infty \leq 1} \left\| \int f(x)\mu(dx) - \frac{1}{N}\sum_{i=1}^N f(x^{(i)}) \right\|_2 \leq \frac{1}{\sqrt{N}}.$$

Proof. Let $\mathcal{F}^N$ be the $\sigma-$algebra generated by $\{x^{(i)}\}_{i=1}^N$ and $\mu(f) := \int f(x)\mu(dx)$. Notice that $\mu(f^2) = \|f\|_2^2$. Then we have that:

$$\mathbb{E}\left[\left(\int f(x)\mu(dx) - \frac{1}{N}\sum_{i=1}^N f(x^{(i)})\right)^2 \mid \mathcal{F}^N\right] = \mathbb{E}\left[\mu(f)^2 - \frac{2}{N}\mu(f)\sum_{i=1}^N f(x^{(i)}) + \frac{1}{N^2}\left(\sum_{i=1}^N f(x^{(i)})\right)^2 \mid \mathcal{F}^N\right],$$
$$= \frac{1}{N^2}\sum_{i=1}^N \mathbb{E}\left[\left(f(x^{(i)})\right)^2 \mid \mathcal{F}^N\right] - \mu(f)^2,$$
$$= \frac{1}{N}\mu(f^2) + \left(\frac{N(N-1)}{N^2} - 1\right)\mu(f)^2,$$
$$= \frac{1}{N}\left(\mu(f^2) - \mu(f)^2\right) \leq \frac{1}{N}\|f\|_\infty^2 < \infty,$$

since $\mu(f)^2 \geq 0$. Note that given $\mathcal{F}^{\mathcal{N}}$, we have that $\mathbb{E}\left[f(x^{(i)})f(x^{(j)}) \mid \mathcal{F}^N\right] = \mathbb{E}\left[f(x^{(i)})\right]\mathbb{E}\left[f(x^{(j)})\right]$ for any $i < j$. Then the result follows from a property of the conditional expectation:

$$\sup_{\|f\|_\infty \leq 1} \left\| \int f(x)\mu(dx) - \frac{1}{N}\sum_{i=1}^N f(x^{(i)}) \right\|_2^2 = \sup_{\|f\|_\infty \leq 1} \mathbb{E}\left[\mathbb{E}\left[\left(\int f(x)\mu(dx) - \frac{1}{N}\sum_{i=1}^N f(x^{(i)})\right)^2 \mid \mathcal{F}^N\right]\right],$$
$$= \sup_{\|f\|_\infty \leq 1} \frac{1}{N}\|f\|_\infty^2 \leq \frac{1}{N}.$$

□

F Importance Sampling

Assume that one needs to obtain sample from $\mu(dx) \in \mathcal{P}(E)$. Let $q(dx) \in \mathcal{P}(E)$ such that μ is absolutely continuous w.r.t. q, and assume that μ and q have densities w.r.t. dx, also denoted by μ

and q. Then the Radon–Nikodym ensures that the following can be well defined:

$$w(x) := \frac{p(x)}{q(x)}.$$

Then for $f \in \mathcal{B}_b(E)$, we have:

$$I := \int_E f(x)p(x)dx = \int_E f(x)w(x)dx.$$

Therefore, to evaluate $\int_E f(x)p(x)dx$, one needs to obtain sample $x^{(i)} \overset{i.i.d.}{\sim} q(x)$, and I can be approximated as follows:

$$\hat{I}_N := \sum_{i=1}^{N} f(x^{(i)})w(x^{(i)}).$$

Indeed, it is clear to see that $\mathbb{E}_q\left[\hat{I}_n\right] = I$ and $\hat{I}_N \to I$ w.p.1. as $N \to \infty$ by the strong law of large numbers, see Geweke (1989) for a detailed asymptotic analysis. Then the variance of $\hat{I}_N$ is given by $\mathbb{V}[\hat{I}_N] = \frac{1}{N}\left(\int_E f(x)^2 w(x)^2 q(x)dx - I^2\right)$. The following characterises the optimal choice of q in terms of $\mathbb{V}[\hat{I}_N]$.

Proposition 47. *Geweke (1989, Theorem 3) The choice* $q(x) = \frac{|f(x)|p(x)}{\int_E |f(x)|p(x)dx}$ *minimises* $\mathbb{V}[\hat{I}_N]$.

Proof. First recall that the Radon–Nikodym ensures that $w(x)$ is always non negative, since p and q are both finite measures. Then the Cauchy–Schwarz inequality gives rise to $\int_E |f(x)|\, w(x)q(x)dx \leq \left(\int_E f(x)^2 w(x)^2 q(x)dx\right)^{1/2}$. Thus we have that $\int_E |f(x)|\, p(x)dx \leq \left(\int_E f(x)^2 w(x)^2 q(x)dx\right)^{1/2}$. Define $\sigma^2 := \int_E f(x)^2 w(x)^2 q(x)dx - I^2$ so that $\left(\int_E f(x)^2 w(x)^2 q(x)dx\right)^{1/2} = \left(\sigma^2 + I^2\right)^{1/2}$, and as a result we obtain:

$$\left(\int_E |f(x)|\, p(x)dx\right)^2 - I^2 \leq \sigma^2.$$

It is clear to see that if there exists a constant $c > 0$ such that $cq(x) = |f(x)|\, p(x)$ then the above inequality becomes equality. Since $\int_E q(x)dx = 1$ by the definition, we obtain $c = \int_E |f(x)|\, p(x)dx$. □

References

Ait-Sahalia, Y. (1996). Testing continuous-time models of the spot interest rate. *The review of financial studies*, 9(2):385–426. 193

Aït-Sahalia, Y. (2002). Maximum likelihood estimation of discretely sampled diffusions: a closed-form approximation approach. *Econometrica*, 70(1):223–262. 165

Ait-Sahalia, Y. (2008). Closed-form likelihood expansions for multivariate diffusions. *The Annals of Statistics*, 36(2):906–937. 170, 177

Aït-Sahalia, Y., Mykland, P. A., and Zhang, L. (2005). How often to sample a continuous-time process in the presence of market microstructure noise. *The review of financial studies*, 18(2):351–416. 101, 165

Ait-Sahalia, Y. and Yu, J. (2008). High frequency market microstructure noise estimates and liquidity measures. Technical report, National Bureau of Economic Research. 101, 165

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723. 127, 128, 130, 131, 148

Allman, E. S., Matias, C., and Rhodes, J. A. (2009). Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, 37(6A):3099–3132. 132

Alòs, E., León, J. A., and Vives, J. (2007). On the short-time behavior of the implied volatility for jump-diffusion models with stochastic volatility. *Finance and Stochastics*, 11(4):571–589. 199

Alquier, P., Friel, N., Everitt, R., and Boland, A. (2016). Noisy monte carlo: Convergence of markov chains with approximate transition kernels. *Statistics and Computing*, 26(1-2):29–47. 224

Andersen, T. G. and Bollerslev, T. (1997). Intraday periodicity and volatility persistence in financial markets. *Journal of empirical finance*, 4(2-3):115–158. 199

Andersen, T. G., Bollerslev, T., Diebold, F. X., and Labys, P. (2001). The distribution of realized exchange rate volatility. *Journal of the American statistical association*, 96(453):42–55. 199

Ando, T. (2007). Bayesian predictive information criterion for the evaluation of hierarchical Bayesian and empirical Bayes models. *Biometrika*, 94(2):443–458. 132

Andrieu, C., Doucet, A., and Holenstein, R. (2010). Particle markov chain monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(3):269–342. 10, 121, 123, 166, 200

Andrieu, C. and Roberts, G. O. (2009). The pseudo-marginal approach for efficient monte carlo computations. *The Annals of Statistics*, 37(2):697–725. 10, 53, 121, 122, 206, 207

Andrieu, C. and Thoms, J. (2008). A tutorial on adaptive mcmc. *Statistics and computing*, 18(4):343–373. 220

Andrieu, C. and Vihola, M. (2015). Convergence properties of pseudo-marginal markov chain monte carlo algorithms. *The Annals of Applied Probability*, 25(2):1030–1077. 54

Azlarov, T. A. and Volodin, N. A. (1982). Laws of large numbers for identically distributed banach-space valued random variables. *Theory of Probability & Its Applications*, 26(3):573–580. 228

Azuma, K. (1967). Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal, Second Series*, 19(3):357–367. 183

Bardenet, R., Doucet, A., and Holmes, C. (2017). On markov chain monte carlo methods for tall data. *The Journal of Machine Learning Research*, 18(1):1515–1557. 224

Barron, A., Birgé, L., and Massart, P. (1999). Risk bounds for model selection via penalization. *Probability theory and related fields*, 113(3):301–413. 155, 162

Basu, A., Harris, I. R., Hjort, N. L., and Jones, M. (1998). Robust and efficient estimation by minimising a density power divergence. *Biometrika*, 85(3):549–559. 224

Bayer, C., Friz, P., and Gatheral, J. (2016). Pricing under rough volatility. *Quantitative Finance*, 16(6):887–904. 199, 200

Beaumont, M. A. (2003). Estimation of population growth or decline in genetically monitored populations. *Genetics*, 164(3):1139–1160. 10, 53

Belfadli, R., Es-Sebaiy, K., and Ouknine, Y. (2011). Parameter estimation for fractional ornstein-uhlenbeck processes: non-ergodic case. *arXiv preprint arXiv:1102.5491*. 200

Bengtsson, T. and Cavanaugh, J. E. (2006). An improved Akaike information criterion for state-space model selection. *Computational Statistics & Data Analysis*, 50(10):2635–2654. 130

Berg, A., Meyer, R., and Yu, J. (2004). Deviance information criterion for comparing stochastic volatility models. *Journal of Business & Economic Statistics*, 22(1):107–120. 132

Besag, J. (1994). Discussion: Markov chains for exploring posterior distributions. *The Annals of Statistics*, 22(4):1734–1741. 10, 55

Beskos, A., Dureau, J., and Kalogeropoulos, K. (2015). Bayesian inference for partially observed stochastic differential equations driven by fractional Brownian motion. *Biometrika*, 102(4):809–827. 200, 210, 216, 217

Beskos, A., Jasra, A., Kantas, N., and Thiery, A. (2016). On the convergence of adaptive sequential monte carlo methods. *The Annals of Applied Probability*, 26(2):1111–1146. 81, 219, 220

Beskos, A., Kalogeropoulos, K., and Pazos, E. (2013a). Advanced MCMC methods for sampling on diffusion pathspace. *Stochastic Processes and their Applications*, 123(4):1415–1453. 10, 66, 67, 68, 208, 216

Beskos, A., Papaspiliopoulos, O., and Roberts, G. (2009). Monte carlo maximum likelihood estimation for discretely observed diffusion processes. *The Annals of Statistics*, 37(1):223–245. 225

Beskos, A., Papaspiliopoulos, O., and Roberts, G. O. (2006). Retrospective exact simulation of diffusion sample paths with applications. *Bernoulli*, 12(6):1077–1098. 165, 167

Beskos, A., Pillai, N., Roberts, G., Sanz-Serna, J.-M., and Stuart, A. (2013b). Optimal tuning of the hybrid Monte Carlo algorithm. *Bernoulli*, 19(5A):1501–1534. 66, 208

Beskos, A., Pinski, F. J., Sanz-Serna, J. M., and Stuart, A. M. (2011). Hybrid monte carlo on hilbert spaces. *Stochastic Processes and their Applications*, 121(10):2201–2230. 10, 66, 68, 71, 208, 209, 220, 225

Betancourt, M. (2017). A conceptual introduction to hamiltonian monte carlo. *arXiv preprint arXiv:1701.02434*. 56, 57, 61

Billingsley, P. (2008). *Probability and measure*. John Wiley & Sons. 229

Bissiri, P. G., Holmes, C. C., and Walker, S. G. (2016). A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):1103–1130. 224

Bollerslev, T. and Jubinski, D. (1999). Equity trading volume and volatility: Latent information arrivals and common long-run dependencies. *Journal of Business & Economic Statistics*, 17(1):9–21. 199

Bou-Rabee, N. and Sanz-Serna, J. M. (2018). Geometric integrators and the hamiltonian monte carlo method. *Acta Numerica*, 27:113–206. 58, 59, 60, 208

Brooks, S., Smith, J., Vehtari, A., Plummer, M., Stone, M., Robert, C. P., Titterington, D., Nelder, J., Atkinson, A., and Dawid, A. (2002). Discussion on the paper by spiegelhalter, best, carlin and van der linde. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 64(4):616–639. 132

Brooks, S. P., Giudici, P., and Roberts, G. O. (2003). Efficient construction of reversible jump markov chain monte carlo proposal distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(1):3–39. 52, 208

Buchholz, A., Chopin, N., and Jacob, P. E. (2018). Adaptive tuning of hamiltonian monte carlo within sequential monte carlo. *arXiv preprint arXiv:1808.07730*. 221

Cappe, O. (2009). Online sequential Monte Carlo EM algorithm. In *2009 IEEE/SP 15th Workshop on Statistical Signal Processing*. 116, 172

Cappé, O., Moulines, E., and Ryden, T. (2005). *Inference in Hidden Markov Models*. Springer Science & Business Media. 99, 103, 116, 135, 136, 165, 185

Carlin, B. P. and Chib, S. (1995). Bayesian model choice via markov chain monte carlo methods. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(3):473–484. 200

Carpenter, J., Clifford, P., and Fearnhead, P. (1999). Improved particle filter for nonlinear problems. *IEE Proceedings-Radar, Sonar and Navigation*, 146(1):2–7. 182

Chapman, D. A. and Pearson, N. D. (2000). Is the short rate drift actually nonlinear? *The Journal of Finance*, 55(1):355–388. 193

Chatterjee, D., Maitra, T., and Bhattacharya, S. (2020). A short note on almost sure convergence of bayes factors in the general set-up. *The American Statistician*, 74(1):17–20. 147

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the american statistical association*, 90(432):1313–1321. 130, 200

Chib, S. and Kuffner, T. A. (2016). Bayes factor consistency. *arXiv preprint arXiv:1607.00292*. 129, 192, 198

Chib, S., Pitt, M. K., and Shephard, N. (2004). Likelihood based inference for diffusion driven models. 175

Chickering, D. M. and Heckerman, D. (1997). Efficient approximations for the marginal likelihood of Bayesian networks with hidden variables. *Machine learning*, 29(2-3):181–212. 133

Chipman, H., George, E. I., McCulloch, R. E., Clyde, M., Foster, D. P., and Stine, R. A. (2001). The practical implementation of bayesian model selection. *Lecture Notes-Monograph Series*, pages 65–134. 206

Chopin, N. (2002). A sequential particle filter method for static models. *Biometrika*, 89(3):539–552. 10, 78, 125, 204, 219

Chopin, N., Jacob, P. E., and Papaspiliopoulos, O. (2013). SMC$\hat{2}$: an efficient algorithm for sequential analysis of state space models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(3):397–426. 10, 121, 125, 126, 166, 200, 219

Chronopoulou, A. and Viens, F. G. (2012). Estimation and pricing under long-memory stochastic volatility. *Annals of Finance*, 8(2-3):379–403. 199

Çınlar, E. (2011). *Probability and stochastics*, volume 261. Springer Science & Business Media. 14, 16

Claeskens, G. and Hjort, N. L. (2008). Model selection and model averaging. *Cambridge Books*. 127, 128, 131, 148, 152, 155

Comte, F., Coutin, L., and Renault, E. (2012). Affine fractional stochastic volatility models. *Annals of Finance*, 8(2-3):337–378. 199

Comte, F. and Renault, E. (1996). Long memory continuous time models. *Journal of Econometrics*, 73(1):101–149. 199

Comte, F. and Renault, E. (1998). Long memory in continuous-time stochastic volatility models. *Mathematical finance*, 8(4):291–323. 199

Cont, R. (2001). Empirical properties of asset returns: stylized facts and statistical issues. 199, 200

Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, 7(2):174–196. 199

Cotter, S. L., Roberts, G. O., Stuart, A. M., and White, D. (2013). Mcmc methods for functions: modifying old algorithms to make them faster. *Statistical Science*, pages 424–446. 71, 225

Craigmile, P. F. (2003). Simulating a class of stationary gaussian processes using the davies–harte algorithm, with application to long memory processes. *Journal of Time Series Analysis*, 24(5):505–511. 213

Crisan, D. and Doucet, A. (2002). A survey of convergence results on particle filtering methods for practitioners. *IEEE Transactions on signal processing*, 50(3):736–746. 110

Crouse, M. S., Nowak, R. D., and Baraniuk, R. G. (1998). Wavelet-based statistical signal processing using hidden markov models. *IEEE Transactions on signal processing*, 46(4):886–902. 99

Csiszár, I. and Shields, P. C. (2000). The consistency of the bic markov order estimator. *The Annals of Statistics*, 28(6):1601–1619. 127

Cuny, C. (2015). A compact lil for martingales in 2-smooth banach spaces with applications. *Bernoulli*, 21(1):374–400. 229

Da Prato, G. (2006). *An introduction to infinite-dimensional analysis*. Springer Science & Business Media. 68, 216, 217

Dahlhaus, R. and Neddermeyer, J. C. (2010). Particle Filter-Based On-Line Estimation of Spot Volatility with Nonlinear Market Microstructure Noise Models. *arXiv preprint arXiv:1006.1860*. 116, 172

Davies, R. B. and Harte, D. (1987). Tests for hurst effect. *Biometrika*, 74(1):95–101. 200, 211

De Gunst, M. and Shcherbakova, O. (2008). Asymptotic behavior of bayes estimators for hidden markov models with application to ion channels. *Mathematical Methods of Statistics*, 17(4):342–356. 115

Dedecker, J. and Merlevède, F. (2008). Convergence rates in the law of large numbers for banach-valued dependent variables. *Theory of Probability & Its Applications*, 52(3):416–438. 229

Del Moral, P. (2004). Feynman-kac formulae. In *Feynman-Kac Formulae*, pages 47–93. Springer. 82, 93, 97, 122, 135

Del Moral, P. (2013). *Mean field simulation for Monte Carlo integration*. Chapman and Hall/CRC. 82, 93

Del Moral, P., Doucet, A., and Jasra, A. (2006). Sequential monte carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):411–436. 10, 78, 80, 81, 200, 204, 224

Del Moral, P., Doucet, A., and Jasra, A. (2012). On adaptive resampling strategies for sequential monte carlo methods. *Bernoulli*, 18(1):252–278. 219

Del Moral, P., Doucet, A., and Singh, S. (2010). Forward smoothing using sequential Monte Carlo. *arXiv preprint arXiv:1012.5390*. 10, 76, 116, 117, 118, 121, 167, 172, 173, 174, 184, 186

Del Moral, P., Doucet, A., and Singh, S. S. (2015). Uniform stability of a particle approximation of the optimal filter derivative. *SIAM Journal on Control and Optimization*, 53(3):1278–1304. 157, 187

Dellaportas, P., Friel, N., and Roberts, G. O. (2006). Bayesian model selection for partially observed diffusion models. *Biometrika*, 93(4):809–825. 52, 177, 193

Delyon, B. and Hu, Y. (2006). Simulation of conditioned diffusion and application to parameter estimation. *Stochastic Processes and their Applications*, 116(11):1660–1675. 175, 176, 179, 197

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38. 116

Deya, A., Neuenkirch, A., and Tindel, S. (2012). A milstein-type scheme without lévy area terms for sdes driven by fractional brownian motion. In *Annales de l'IHP Probabilités et statistiques*, volume 48, pages 518–550. 215

Diel, R., Le Corff, S., and Lerasle, M. (2020). Learning the distribution of latent variables in paired comparison models with round-robin scheduling. 162

Ding, J., Tarokh, V., and Yang, Y. (2017). Bridging AIC and BIC: a new criterion for autoregression. *IEEE Transactions on Information Theory*, 64(6):4024–4043. 155

Ding, J., Tarokh, V., and Yang, Y. (2018). Model selection techniques: An overview. *IEEE Signal Processing Magazine*, 35(6):16–34. 155

Ding, Z., Granger, C. W., and Engle, R. F. (1993). A long memory property of stock market returns and a new model. *Journal of empirical finance*, 1(1):83–106. 199

Douc, R. and Cappé, O. (2005). Comparison of resampling schemes for particle filtering. In *Image and Signal Processing and Analysis, 2005. ISPA 2005. Proceedings of the 4th International Symposium on*, pages 64–69. IEEE. 75, 182

Douc, R., Garivier, A., Moulines, E., and Olsson, J. (2011a). Sequential monte carlo smoothing for general state space hidden markov models. *The Annals of Applied Probability*, 21(6):2109–2145. 184

Douc, R. and Moulines, E. (2012). Asymptotic properties of the maximum likelihood estimation in misspecified hidden markov models. *The Annals of Statistics*, 40(5):2697–2732. 115, 162

Douc, R., Moulines, E., Olsson, J., and Van Handel, R. (2011b). Consistency of the maximum likelihood estimator for general hidden Markov models. *the Annals of Statistics*, 39(1):474–513. 115, 162

Douc, R., Moulines, E., Priouret, P., and Soulier, P. (2018). *Markov chains*. Springer. 25, 27, 28

Douc, R., Moulines, E., and Rydén, T. (2004). Asymptotic properties of the maximum likelihood estimator in autoregressive models with markov regime. *The Annals of statistics*, 32(5):2254–2304. 115, 132, 136, 137, 225

Douc, R., Moulines, E., and Stoffer, D. (2014). *Nonlinear time series: Theory, methods and applications with R examples*. Chapman and Hall/CRC. 72, 99, 115, 127, 132, 134, 136, 137, 139, 140, 142, 165

Douc, R., Olsson, J., and Roueff, F. (2016a). Posterior consistency for partially observed markov models. *arXiv preprint arXiv:1608.06851*. 115

Douc, R., Olsson, J., and Roueff, F. (2020). Posterior consistency for partially observed Markov models. *Stochastic Processes and their Applications*, 130(2):733–759. 162

Douc, R., Roueff, F., and Sim, T. (2016b). The maximizing set of the asymptotic normalized log-likelihood for partially observed Markov chains. *The Annals of Applied Probability*, 26(4):2357–2383. 162

Doucet, A., Godsill, S., and Andrieu, C. (2000). On sequential Monte Carlo sampling methods for bayesian filtering. *Statistics and computing*, 10(3):197–208. 117, 173

Doucet, A. and Johansen, A. M. (2009). A tutorial on particle filtering and smoothing: Fifteen years later. *Handbook of nonlinear filtering*, 12(656-704):3. 72, 80, 99, 157, 204, 220

Duane, S., Kennedy, A. D., Pendleton, B. J., and Roweth, D. (1987). Hybrid monte carlo. *Physics letters B*, 195(2):216–222. 10, 62, 200, 208

Durbin, J. and Koopman, S. J. (2012). *Time series analysis by state space methods*, volume 38. Oxford University Press. 100

Durham, G. B. (2003). Likelihood-based specification analysis of continuous-time models of the short-term interest rate. *Journal of Financial Economics*, 70(3):463–487. 165, 193

Durham, G. B. and Gallant, A. R. (2002). Numerical techniques for maximum likelihood estimation of continuous-time diffusion processes. *Journal of Business & Economic Statistics*, 20(3):297–338. 197

Durmus, A., Moulines, E., and Saksman, E. (2017). On the convergence of hamiltonian monte carlo. *arXiv preprint arXiv:1705.00166*. 66

Eberle, A. (2020). Markov processes. 30, 40

Eguchi, S. and Masuda, H. (2018). Schwarz type model comparison for laq models. *Bernoulli*, 24(3):2278–2327. 192

El Euch, O. and Rosenbaum, M. (2018). Perfect hedging in rough heston models. *The Annals of Applied Probability*, 28(6):3813–3856. 199

El Euch, O. and Rosenbaum, M. (2019). The characteristic function of rough heston models. *Mathematical Finance*, 29(1):3–38. 199

Eraker, B., Johannes, M., and Polson, N. (2003). The impact of jumps in volatility and returns. *The Journal of Finance*, 58(3):1269–1300. 191, 192

Everitt, R. G., Johansen, A. M., Rowing, E., and Evdemon-Hogan, M. (2017). Bayesian model comparison with un-normalised likelihoods. *Statistics and Computing*, 27(2):403–422. 200

Fearnhead, P., Papaspiliopoulos, O., and Roberts, G. O. (2008). Particle filters for partially observed diffusions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(4):755–777. 167, 174, 177

Fearnhead, P. and Taylor, B. M. (2013). An adaptive sequential monte carlo sampler. *Bayesian analysis*, 8(2):411–438. 220, 221

Feng, C., Wang, H., Chen, T., and Tu, X. M. (2014). On exact forms of taylorâs theorem for vector-valued functions. *Biometrika*, 101(4):1003–1003. 226

Forde, M. and Zhang, H. (2017). Asymptotics for rough stochastic volatility models. *SIAM Journal on Financial Mathematics*, 8(1):114–145. 199

Friel, N., McKeone, J., Oates, C. J., and Pettitt, A. N. (2017). Investigation of the widely applicable Bayesian information criterion. *Statistics and Computing*, 27(3):833–844. 133, 134

Fuchs, C. (2013). *Inference for diffusion processes: with applications in life sciences*. Springer Science & Business Media. 168

Fukasawa, M. (2017). Short-time at-the-money skew and rough fractional volatility. *Quantitative Finance*, 17(2):189–198. 199

Gal, Y. (2016). Uncertainty in deep learning. 224

Gales, M. and Young, S. (2008). The application of hidden markov models in speech recognition. *Foundations and Trends® in Signal Processing*, 1(3):195–304. 99

Gassiat, E. and Boucheron, S. (2003). Optimal error exponents in hidden Markov models order estimation. *IEEE Transactions on Information Theory*, 49(4):964–980. 127, 152

Gassiat, E. and Rousseau, J. (2014). About the posterior distribution in hidden markov models with unknown number of states. *Bernoulli*, 20(4):2039–2075. 115

Gatheral, J., Jaisson, T., and Rosenbaum, M. (2018). Volatility is rough. *Quantitative Finance*, 18(6):933–949. 199, 200

Gelfand, A. E. and Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, 85(410):398–409. 200

Gelman, A., Hwang, J., and Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and computing*, 24(6):997–1016. 128, 131, 133

Gelman, A. and Meng, X.-L. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical science*, pages 163–185. 130, 203

Geweke, J. (1989). Bayesian inference in econometric models using monte carlo integration. *Econometrica: Journal of the Econometric Society*, pages 1317–1339. 231

Geyer, C. J. (1994). On the convergence of monte carlo maximum likelihood calculations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(1):261–274. 225

Gloaguen, P., Étienne, M.-P., and Le Corff, S. (2018). Online sequential monte carlo smoother for partially observed diffusion processes. *EURASIP Journal on Advances in Signal Processing*, 2018(1):9. 167

Gloter, A. and Hoffmann, M. (2007). Estimation of the hurst parameter from discrete noisy data. *The Annals of Statistics*, 35(5):1947–1974. 199

Golightly, A. and Wilkinson, D. J. (2008). Bayesian inference for nonlinear multivariate diffusion models observed with error. *Computational Statistics & Data Analysis*, 52(3):1674–1693. 175, 200

Golub, G. H. and Van Loan, C. F. (2012). *Matrix computations*, volume 3. JHU Press. 212

Gonçalves, F. B. and Roberts, G. O. (2014). Exact simulation problems for jump-diffusions. *Methodology and Computing in Applied Probability*, 16(4):907–930. 178

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. MIT press. 187

Gordon, N. J., Salmond, D. J., and Smith, A. F. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. In *IEE Proceedings F (Radar and Signal Processing)*, volume 140, pages 107–113. IET. 10, 107, 109, 118, 172

Green, P. J. (1995). Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82(4):711–732. 10, 50, 51, 52, 203, 208

Green, P. J. and Richardson, S. (2002). Hidden markov models and disease mapping. *Journal of the American statistical association*, 97(460):1055–1070. 99

Guarniero, P., Johansen, A. M., and Lee, A. (2017). The iterated auxiliary particle filter. *Journal of the American Statistical Association*, 112(520):1636–1647. 10, 113, 114

Guennoun, H., Jacquier, A., Roome, P., and Shi, F. (2018). Asymptotic behavior of the fractional heston model. *SIAM Journal on Financial Mathematics*, 9(3):1017–1045. 199

Hairer, M. (2010). Convergence of markov processes. 21

Hairer, M. (2018). Ergodic properties of markov processes. 20, 25, 27

Hairer, M. and Mattingly, J. C. (2011). Yet another look at harrisâ ergodic theorem for markov chains. In *Seminar on Stochastic Analysis, Random Fields and Applications VI*, pages 109–117. Springer. 25, 30

Hannan, E. J. and Quinn, B. G. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 190–195. 156

Hansen, L. P. and Scheinkman, J. A. (1993). Back to the future: generating moment implications for continuous-time markov processes. 165

Hansen, P. R. and Lunde, A. (2006). Realized variance and market microstructure noise. *Journal of Business & Economic Statistics*, 24(2):127–161. 101, 165

Hastie, D. I. and Green, P. J. (2012). Model choice using reversible jump markov chain monte carlo. *Statistica Neerlandica*, 66(3):309–338. 50

Heston, S. L. (1993). A closed-form solution for options with stochastic volatility with applications to bond and currency options. *The review of financial studies*, 6(2):327–343. 190

Higham, D. J. (2001). An algorithmic introduction to numerical simulation of stochastic differential equations. *SIAM review*, 43(3):525–546. 40

Hoffman, M. D. and Gelman, A. (2014). The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, 15(1):1593–1623. 225

Hoffmann-Jorgensen, J. and Pisier, G. (1976). The law of large numbers and the central limit theorem in banach spaces. *The Annals of Probability*, 4(4):587–599. 228, 229

Holmes, C. and Walker, S. (2017). Assigning a value to a power likelihood in a general bayesian model. *Biometrika*, 104(2):497–503. 224

Hu, Y. and Nualart, D. (2010). Parameter estimation for fractional ornstein–uhlenbeck processes. *Statistics & probability letters*, 80(11-12):1030–1038. 200

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 221–233. University of California Press. 130

Hull, J. and White, A. (1987). The pricing of options on assets with stochastic volatilities. *The journal of finance*, 42(2):281–300. 100

Hurvich, C. M. and Tsai, C. (1990). The impact of model selection on inference in linear regression. *The American Statistician*, 44(3):214–217. 127

Hurvich, C. M. and Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, 76(2):297–307. 127

Ibragimov, R. and Sharakhmetov, S. (1999). Analogues of khintchine, marcinkiewicz–zygmund and rosenthal inequalities for symmetric statistics. *Scandinavian journal of statistics*, 26(4):621–633. 149

Jasra, A., Stephens, D. A., Doucet, A., and Tsagaris, T. (2011). Inference for lévy-driven stochastic volatility models via adaptive sequential monte carlo. *Scandinavian Journal of Statistics*, 38(1):1–22. 219, 220

Jeffreys, H. (1998). *The theory of probability*. OUP Oxford. 127, 129, 198, 202

Johannes, M. S., Polson, N. G., and Stroud, J. R. (2009). Optimal filtering of jump diffusions: Extracting latent states from asset prices. *The Review of Financial Studies*, 22(7):2759–2799. 191, 192

Johansen, A. M. and Doucet, A. (2008). A note on auxiliary particle filters. *Statistics & Probability Letters*, 78(12):1498–1504. 10, 111, 113, 182

Jones, C. S. (2003). Nonlinear mean reversion in the short-term interest rate. *The Review of Financial Studies*, 16(3):793–843. 193

Kalogeropoulos, K., Roberts, G. O., and Dellaportas, P. (2010). Inference for stochastic volatility models using time change transformations. *The Annals of Statistics*, 38(2):784–807. 170, 172, 177, 200

Kantas, N., Doucet, A., Singh, S. S., Maciejowski, J., and Chopin, N. (2015). On particle methods for parameter estimation in state-space models. *Statistical science*, 30(3):328–351. 76, 117, 172

Karagiannis, G. and Andrieu, C. (2013). Annealed importance sampling reversible jump mcmc algorithms. *Journal of Computational and Graphical Statistics*, 22(3):623–648. 206, 208

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the american statistical association*, 90(430):773–795. 127, 129, 192, 198, 202, 203, 206

Kass, R. E., Tierney, L., and Kadane, J. B. (1990). The validity of posterior expansions based on Laplace's method. In *In Bayesian and Likelihood Methods in Statistics and Econometrics, edited by S. Geisser, J. S. Hodges, S. J. Press and A. Zellner*, pages 473–488. North-Holland Amsterdam. 130, 144, 145

Kass, R. E., Tierney, L., and Kadane, J. B. (1991). Laplaces method in bayesian analysis. *Contemporary Mathematics*, 115:89–99. 203

Kessler, M. (1997). Estimation of an ergodic diffusion from discrete observations. *Scandinavian Journal of Statistics*, 24(2):211–229. 165

Kessler, M., Lindner, A., and Sorensen, M. (2012). *Statistical methods for stochastic differential equations*. Chapman and Hall/CRC. 165

Kessler, M. and Sørensen, M. (1999). Estimating equations based on eigenfunctions for a discretely observed diffusion process. *Bernoulli*, 5(2):299–314. 165

Khan, M. E., Nielsen, D., Tangkaratt, V., Lin, W., Gal, Y., and Srivastava, A. (2018). Fast and scalable bayesian deep learning by weight-perturbation in adam. *arXiv preprint arXiv:1806.04854*. 224

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*. 186, 187

Kitagawa, G. (1987). Non-gaussian stateï¿œSpace modeling of nonstationary time series. *Journal of the American statistical association*, 82(400):1032–1041. 117, 173

Kitagawa, G. (1996). Monte Carlo filter and smoother for non-Gaussian nonlinear state space models. *Journal of computational and graphical statistics*, 5(1):1–25. 10, 75, 182

Kloeden, P. E. and Platen, E. (2013). *Numerical solution of stochastic differential equations*, volume 23. Springer Science & Business Media. 166

Kong, A. (1992). A note on importance sampling using standardized weights. 76

Kong, A., Liu, J. S., and Wong, W. H. (1994). Sequential imputations and bayesian missing data problems. *Journal of the American statistical association*, 89(425):278–288. 10, 73, 76, 220

Konishi, S. and Kitagawa, G. (1996). Generalised information criteria in model selection. *Biometrika*, 83(4):875–890. 127

Kou, S. C. (2008). Stochastic modeling in nanoscale biophysics: subdiffusion within proteins. *The Annals of Applied Statistics*, 2(2):501–535. 199

Kreyszig, E. (1978). *Introductory functional analysis with applications*, volume 1. wiley New York. 24, 25

Lang, S. (2012). *Real and functional analysis*, volume 142. Springer Science & Business Media. 144, 226

Le Corff, S., Fort, G., et al. (2013). Online expectation maximization based algorithms for inference in hidden markov models. *Electronic Journal of Statistics*, 7:763–792. 162

Le Gland, F. and Mevel, L. (1997). Asymptotic behaviour of the mle in hidden markov models. In *Proceedings of the 4th European Control Conference, Bruxelles 1997*. 158, 184

LeGland, F. and Mevel, L. (1997). Recursive estimation in hidden markov models. In *Decision and Control, 1997., Proceedings of the 36th IEEE Conference on*, volume 4, pages 3468–3473. IEEE. 118, 119, 185

Lehéricy, L. (2018). Nonasymptotic control of the MLE for misspecified nonparametric hidden Markov models. *arXiv preprint arXiv:1807.03997*. 162

Liu, J. S. and Chen, R. (1998). Sequential monte carlo methods for dynamic systems. *Journal of the American statistical association*, 93(443):1032–1044. 76

Livingstone, S., Betancourt, M., Byrne, S., and Girolami, M. (2019). On the geometric ergodicity of hamiltonian monte carlo. *Bernoulli*, 25(4A):3109–3138. 66

Lobato, I. N. and Savin, N. E. (1998). Real and spurious long-memory properties of stock-market data. *Journal of Business & Economic Statistics*, 16(3):261–268. 199

Lobato, I. N. and Velasco, C. (2000). Long memory in stock-market trading volume. *Journal of Business & Economic Statistics*, 18(4):410–427. 199

Louizos, C., Shalit, U., Mooij, J. M., Sontag, D., Zemel, R., and Welling, M. (2017). Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems*, pages 6446–6456. 224

Lyne, A.-M., Girolami, M., Atchadé, Y., Strathmann, H., and Simpson, D. (2015). On russian roulette estimates for bayesian inference with doubly-intractable likelihoods. *Statistical science*, 30(4):443–467. 200

Lysy, M. and Pillai, N. S. (2013). Statistical inference for stochastic differential equations with memory. *arXiv preprint arXiv:1307.1164*. 214

Malik, S. and Pitt, M. K. (2011). Particle filters for continuous likelihood evaluation and maximisation. *Journal of Econometrics*, 165(2):190–209. 225

Mamon, R. S. and Elliott, R. J. (2007). *Hidden markov models in finance*, volume 460. Springer. 99

Mandelbrot, B. B. and Van Ness, J. W. (1968). Fractional brownian motions, fractional noises and applications. *SIAM review*, 10(4):422–437. 199

Mao, X. (2007). *Stochastic differential equations and applications*. Elsevier. 40

McLeish, D. L. (1974). Dependent central limit theorems and invariance principles. *the Annals of Probability*, 2(4):620–628. 141

Mengersen, K. L. and Tweedie, R. L. (1996). Rates of convergence of the hastings and metropolis algorithms. *The annals of Statistics*, 24(1):101–121. 48, 49

Mononen, T. (2015). A case study of the widely applicable Bayesian information criterion and its optimality. *Statistics and Computing*, 25(5):929–940. 134

Mourier, E. (1953). Eléments aléatoires dans un espace de banach. In *Annales de l'institut Henri Poincaré*, volume 13, pages 161–244. 228

Naesseth, C. A., Lindsten, F., and Schön, T. B. (2019). Elements of sequential monte carlo. *arXiv preprint arXiv:1903.04797*. 72

Neal, R. M. (2001). Annealed importance sampling. *Statistics and computing*, 11(2):125–139. 78, 79, 125, 134, 204

Neal, R. M. (2011). MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2(11):2. 57, 208, 216, 220

Neuenkirch, A. and Tindel, S. (2014). A least square-type procedure for parameter estimation in stochastic differential equations with additive fractional noise. *Statistical Inference for Stochastic Processes*, 17(1):99–120. 200

Neuenkirch, A., Tindel, S., and Unterberger, J. (2010). Discretizing the fractional lévy area. *Stochastic Processes and their Applications*, 120(2):223–254. 215

Newton, M. A. and Raftery, A. E. (1994). Approximate bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(1):3–26. 203

Nishii, R. (1988). Maximum likelihood principle and model selection when the true model is unspecified. *Journal of Multivariate analysis*, 27(2):392–403. 131, 152, 155, 192

Øksendal, B. (2003). Stochastic differential equations. In *Stochastic differential equations*, pages 65–84. Springer. 31, 32, 33, 34, 37, 101, 170, 171

Øksendal, B. and Sulem, A. (2007). *Applied Stochastic Control of Jump Diffusions*. Springer Science & Business Media. 165

Olsson, J. and Alenlöv, J. W. (2020). Particle-based online estimation of tangent filters with application to parameter estimation in nonlinear state-space models. *Annals of the Institute of Statistical Mathematics*, 72(2):545–576. 163

Olsson, J. and Westerborn, J. (2017). Efficient particle-based online smoothing in general hidden markov models: the paris algorithm. *Bernoulli*, 23(3):1951–1996. 183

Papaspiliopoulos, O. and Roberts, G. O. (2009). Importance sampling techniques for estimation of diffusions models. 176, 197

Papaspiliopoulos, O., Roberts, G. O., and Stramer, O. (2013). Data augmentation for diffusions. *Journal of Computational and Graphical Statistics*, 22(3):665–688. 176

Pavliotis, G. A. (2014). *Stochastic processes and applications: diffusion processes, the Fokker-Planck and Langevin equations*, volume 60. Springer. 14, 20, 31, 38, 39

Perrin, E., Harba, R., Jennane, R., and Iribarren, I. (2002). Fast and exact synthesis for 1-d fractional brownian motion and fractional gaussian noises. *IEEE Signal Processing Letters*, 9(11):382–384. 213

Pitt, M. K., Malik, S., and Doucet, A. (2014). Simulated likelihood inference for stochastic volatility models using continuous particle filtering. *Annals of the Institute of Statistical Mathematics*, 66(3):527–552. 158

Pitt, M. K. and Shephard, N. (1999). Filtering via simulation: Auxiliary particle filters. *Journal of the American statistical association*, 94(446):590–599. 10, 111, 113, 182

Platen, E. (1999). An introduction to numerical methods for stochastic differential equations. *Acta numerica*, 8:197–246. 40

Pouzo, D., Psaradakis, Z., and Sola, M. (2016). Maximum Likelihood Estimation in Possibly Misspecified Dynamic Models with Time-Inhomogeneous Markov Regimes. *arXiv preprint arXiv:1612.04932*. 163

Poyiadjis, G., Doucet, A., and Singh, S. S. (2011). Particle approximations of the score and observed information matrix in state space models with application to parameter estimation. *Biometrika*, 98(1):65–80. 10, 118, 119, 156, 157, 158, 159, 167, 184, 185

Quiroz, M., Kohn, R., Villani, M., and Tran, M.-N. (2019). Speeding up mcmc by efficient data subsampling. *Journal of the American Statistical Association*, 114(526):831–843. 224

Rangapuram, S. S., Seeger, M. W., Gasthaus, J., Stella, L., Wang, Y., and Januschowski, T. (2018). Deep state space models for time series forecasting. In *Advances in neural information processing systems*, pages 7785–7794. 224

Rao, B. P. (2011). *Statistical inference for fractional diffusion processes*. John Wiley & Sons. 199, 200, 210

Rebeschini, P. and Van Handel, R. (2015). Can local particle filters beat the curse of dimensionality? *The Annals of Applied Probability*, 25(5):2809–2866. 94

Reddi, S. J., Kale, S., and Kumar, S. (2019). On the convergence of adam and beyond. *arXiv preprint arXiv:1904.09237*. 187

Robert, C. (2007). *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Science & Business Media. 206

Robert, C. and Casella, G. (2013). *Monte Carlo statistical methods*. Springer Science & Business Media. 42, 51

Roberts, G. O., Gelman, A., and Gilks, W. R. (1997). Weak convergence and optimal scaling of random walk metropolis algorithms. *The annals of applied probability*, 7(1):110–120. 49, 50

Roberts, G. O. and Rosenthal, J. S. (1998). Optimal scaling of discrete approximations to langevin diffusions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(1):255–268. 55

Roberts, G. O. and Rosenthal, J. S. (2004). General state space markov chains and mcmc algorithms. *Probability surveys*, 1:20–71. 42

Roberts, G. O. and Stramer, O. (2001). On inference for partially observed nonlinear diffusion models using the Metropolis–Hastings algorithm. *Biometrika*, 88(3):603–621. 169, 170, 172, 175, 200

Roberts, G. O. and Tweedie, R. L. (1996a). Exponential convergence of langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363. 10, 55

Roberts, G. O. and Tweedie, R. L. (1996b). Geometric convergence and central limit theorems for multidimensional hastings and metropolis algorithms. *Biometrika*, 83(1):95–110. 47, 48, 49, 55

Sanz-Serna, J. (2014). Markov chain monte carlo and numerical differential equations. In *Current challenges in stability issues for numerical differential equations*, pages 39–88. Springer. 58, 62

Särkkä, S. and Sottinen, T. (2008). Application of Girsanov theorem to particle filtering of discretely observed continuous-time non-linear systems. *Bayesian Analysis*, 3(3):555–584. 167, 174

Schauer, M., Van Der Meulen, F., and Van Zanten, H. (2017). Guided proposals for simulating multi-dimensional diffusion bridges. *Bernoulli*, 23(4A):2917–2950. 197

Schervish, M. J. (2012). *Theory of statistics*. Springer Science & Business Media. 144, 145

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464. 127, 128, 130, 144, 192

Shao, J. (1997). An asymptotic theory for linear model selection. *Statistica Sinica*, pages 221–242. 155

Shao, S., Jacob, P. E., Ding, J., and Tarokh, V. (2018). Bayesian model comparison with the hyvarinen score: computation and consistency. *Journal of the American Statistical Association*, (just-accepted):1–33. 127

Shephard, N. and Andersen, T. G. (2009). Stochastic volatility: origins and overview. In *Handbook of financial time series*, pages 233–254. Springer. 100

Sherlock, C. (2018). Reversible markov chains: variational representations and ordering. *arXiv preprint arXiv:1809.01903*. 46

Sherlock, C. and Roberts, G. (2009). Optimal scaling of the random walk metropolis on elliptically symmetric unimodal targets. *Bernoulli*, 15(3):774–798. 220

Shibata, R. (1976). Selection of the order of an autoregressive model by akaike's information criterion. *Biometrika*, 63(1):117–126. 127

Shibata, R. (1980). Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *The Annals of statistics*, pages 147–164. 155

Shibata, R. (1981). An optimal selection of regression variables. *Biometrika*, 68(1):45–54. 155

Sin, C.-Y. and White, H. (1996). Information criteria for selecting possibly misspecified parametric models. *Journal of Econometrics*, 71(1-2):207–225. 155, 192

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639. 128, 131, 132

Stanton, R. (1997). A nonparametric model of term structure dynamics and the market price of interest rate risk. *The Journal of Finance*, 52(5):1973–2002. 193

Stoltz, G. and Rousset, M. (2010). *Free energy computations: A mathematical perspective*. World Scientific. 56, 59, 62

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):44–47. 155

Stout, W. F. (1970). A martingale analogue of kolmogorov's law of the iterated logarithm. *Probability Theory and Related Fields*, 15(4):279–290. 127, 143

Ströjby, J. and Olsson, J. (2009). Efficient particle-based likelihood estimation in partially observed diffusion processes. *IFAC Proceedings Volumes*, 42(10):1364–1369. 167

Sussmann, H. J. (1978). On the gap between deterministic and stochastic ordinary differential equations. *The Annals of Probability*, pages 19–41. 213

Sutskever, I., Martens, J., Dahl, G., and Hinton, G. (2013). On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pages 1139–1147. 197

Tadic, V. B. (2010). Analyticity, convergence, and convergence rate of recursive maximum-likelihood estimation in hidden markov models. *IEEE Transactions on Information Theory*, 56(12):6406–6432. 158

Tadic, V. Z. and Doucet, A. (2018). Asymptotic properties of recursive maximum likelihood estimation in non-linear state-space models. *arXiv preprint arXiv:1806.09571*. 119, 146, 185

Takeuchi, K. (1976). Distribution of information statistics and validity criteria of models. mathematical science, 153: 12-18, 1976. 130, 148

Taylor, S. J. (1982). Financial returns modelled by the product of two stochastic processes-a study of the daily sugar prices 1961-75. *Time series analysis: theory and practice*, 1:203–226. 100

Tierney, L. (1994). Markov chains for exploring posterior distributions. *the Annals of Statistics*, pages 1701–1728. 42

Tierney, L. (1998). A note on metropolis-hastings kernels for general state spaces. *Annals of applied probability*, pages 1–9. 42, 43, 44, 46, 51

van der Meulen, F. and Schauer, M. (2017). Bayesian estimation of discretely observed multi-dimensional diffusion processes using guided proposals. *Electronic Journal of Statistics*, 11(1):2358–2396. 197

Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press. 226

Varin, C. and Vidoni, P. (2005). A note on composite likelihood inference and model selection. *Biometrika*, 92(3):519–528. 127

Vergé, C., Dubarry, C., Del Moral, P., and Moulines, E. (2015). On parallel implementation of sequential monte carlo methods: the island particle model. *Statistics and Computing*, 25(2):243–260. 93

Wagner, W. (1989). Undiased Monte Carlo estimators for functionals of weak solutions of stochastic differential equations. *Stochastics: An International Journal of Probability and Stochastic Processes*, 28(1):1–20. 165

Watanabe, S. (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11(Dec):3571–3594. 128, 132, 133

Watanabe, S. (2013). A widely applicable Bayesian information criterion. *Journal of Machine Learning Research*, 14(Mar):867–897. 128, 133, 134

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the Econometric Society*, pages 1–25. 130

Wood, A. T. and Chan, G. (1994). Simulation of stationary gaussian processes in [0, 1] d. *Journal of computational and graphical statistics*, 3(4):409–432. 10, 211, 212, 213

Xiao, W. and Yu, J. (2019). Asymptotic theory for estimating drift parameters in the fractional vasicek model. *Econometric Theory*, 35(1):198–231. 200

Xiao, W., Zhang, W., and Xu, W. (2011). Parameter estimation for fractional ornstein–uhlenbeck processes at discrete observation. *Applied Mathematical Modelling*, 35(9):4196–4207. 199, 200

Yang, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation. *Biometrika*, 92(4):937–950. 155

Yonekura, S., Beskos, A., and Singh, S. (2018). Asymptotic analysis of model selection criteria for general hidden markov models. *arXiv preprint arXiv:1811.11834*. 192

Yoon, B.-J. (2009). Hidden markov models and their applications in biological sequence analysis. *Current genomics*, 10(6):402–415. 99

Yu, J. and Meyer, R. (2006). Multivariate stochastic volatility models: Bayesian estimation and model comparison. *Econometric Reviews*, 25(2-3):361–384. 132

Yuan, K., Ying, B., and Sayed, A. H. (2016). On the influence of momentum acceleration on online learning. *The Journal of Machine Learning Research*, 17(1):6602–6667. 197

Zhou, Y., Johansen, A. M., and Aston, J. A. (2016). Toward automatic model comparison: an adaptive sequential Monte Carlo approach. *Journal of Computational and Graphical Statistics*, 25(3):701–726. 130, 202, 204, 208